nature

www.nature.com/nature

Vol 460 | Issue no. 7256 | 6 August 2009

A model approach

More development work is needed to help computer simulations inform economic policy.

Models are everywhere in economics. They range from the pencil-and-paper equations used for academic analyses of market behaviour, to the computer forecasts used by central banks, such as the Bank of England and the US Federal Reserve System, to determine the likely effects of interest-rate adjustments.

But the reputation of economic models has been tarnished of late. Virtually none anticipated the global financial meltdown that began two years ago this summer (see pages 680 and 685). The finger-pointing seems likely to go on indefinitely: were the models flawed? Or were policy-makers at fault for ignoring the warnings?

What is clear is that economic models need to improve. The ability to run policy options through a believable set of 'what-if' scenarios could be useful to forestall future economic crises, and to inform debate, such as that over the labyrinthine efforts to reform the US health-care system.

The field could benefit from lessons learned in the large-scale modelling of other complex phenomena, such as climate change and epidemics (see page 687). Those lessons, taken together with lessons from the downturn, suggest an ambitious research agenda — not just for economists, but for psychologists, political and social scientists, computer researchers and more.

First, details matter. Government regulators rely on dynamic stochastic general equilibrium (DSGE) simulations, which can make sophisticated extrapolations of past economic data. But these models do little to incorporate information about the financial sector, which is where the current crisis began. Which company was entering into what kind of arrangements with another, for example, and how were they all interconnected? And most models don't even attempt to incorporate the psychological insights gained from behavioural economics, and so ignore shifting attitudes towards risk, and the spread of fear — both major contributors to the crisis. The comparatively few modelling efforts that do try to include these factors deserve support — and many more such efforts are needed.

Second, models should evolve through vigorous competition. As the articles in this issue show, advocates of agent-based modelling techniques, which represent each individual or company with an 'agent', claim that their programs can often account for economic phenomena much better than can DSGE simulations. Such claims need to be addressed empirically. The economics community should try to agree on a standard set of test cases analogous to those used by climate modellers, whose challenges can include being able to reproduce El Niño oscillations. Economic modellers should also consider adopting the modular architecture used in many climate models. This approach makes it easy to aggregate smaller models into more comprehensive simulations, while still allowing steady improvement in each piece. A sub-model for ocean circulation, say, can be switched for an alternative circulation module without changing anything else.

Third, modellers seeking to make a real difference in the world should concentrate on the tangible, immediate questions that decision-makers actually worry about. A good example to follow is that of pandemic planning, in which simulations are already in widespread use to help officials decide when to close schools and other public gathering places, and how best to mount a vaccination campaign. The simulations alone cannot answer such questions, nor can they replace judgement. But by helping officials frame the problem, organize the available information and identify which factors matter, they can make judgements better informed. ■

Science under attack

Congress should stop playing politics with the peer-review process.

In a depressingly familiar display of irresponsible politicking, the US House of Representatives has taken aim at three studies funded by the National Institutes of Health (NIH). Representative Darrell Issa (Republican, California) introduced an amendment killing the projects on 24 July, during a debate on the NIH's 2010 budget. The House passed the amendment by a voice vote.

Issa was unhappy that the studies looked at substance abuse and HIV risk behaviour, and that the subjects were outside the United States. One focused on Russian alcoholics, another on female sex workers in China and a third on female and transgender prostitutes in Thailand. All three passed muster with NIH peer reviewers, and together would cost about $5 million over five years. Issa wanted that money to be spent at home, and complained that HIV had been heavily studied already. But his reasoning is specious: alcoholism, prostitution and HIV do not respect borders, and any behavioural information that could help slow the transmission of HIV is crucial. Some 33 million people are infected worldwide, and a vaccine is nowhere in sight.

Issa's tactic is not new. Since 2003, conservative House Republicans have tried at least five times to strip funding from peer-reviewed projects that drew their ire. Such meddling threatens to undermine the peer-review process as well as potentially eroding the public's trust that science is above politics.

Also worrying is the House Democrats' acquiescence to Issa's amendment. Democrats facing tough re-election bids hoped to dodge Republican attacks in media adverts in their home districts that might have resulted from opposing Issa. Their assumption is that the amendment can be quietly removed when House and Senate negotiators meet to square their versions of the NIH bill before a final vote on it. But Congress should renounce all tactics that undermine peer review — and cease indulging those who use them. ■

© 2009 Macmillan Publishers Limited. All rights reserved

RESEARCH HIGHLIGHTS

EXOPLANETS

Avoiding shrinkage

Astrophys. J. **700,** 1921–1932 (2009)

Many of the planets discovered outside the Solar System are bigger than Jupiter. Some are larger than expected given the steady shrinking that occurs as gas-giant planets cool.

Using simulations, Laurent Ibgui and Adam Burrows of Princeton University in New Jersey have shown how the enormous size of some of these planets can be attributed to peculiar conditions at their birth. Just after formation, the planet could wind up in an eccentric orbit close to its star. Later, tides between the star and planet would pump heat into the planet's interior and inflate it.

The process, the authors say, might explain the lack of shrinkage for some older giants, giving them the look of a planet a billion years younger.

CANCER BIOLOGY

HPV's unexpected effect

Cancer Prev. Res. doi:10.1158/1940-6207.capr-09-0149 (2009)

People infected with human papillomavirus (HPV) have a better chance of surviving a type of head and neck cancer than those without the infection. The findings may help explain why black cancer patients fare worse than whites.

Kevin Cullen of the University of Maryland in Baltimore and his colleagues found that whites with squamous cell carcinoma of the throat survived about three times longer than blacks with this condition. By analysing biopsy specimens from 196 whites and 28 blacks, the authors determined that this disparity might be explained by HPV status: the survival rate was two-and-a-half times higher for infected patients than uninfected patients, and white patients were almost nine times more likely to be HPV-positive than blacks.

Cullen says HPV may make tumours more vulnerable to chemotherapy and radiation.

MATERIALS SCIENCE

Foam finesse

Colloids Surf. A: Physicochem. Eng. Aspects doi:10.1016/j.colsurfa.2009.05.010 (2009)

When gas rushes through solidifying foam to create porous polymers — used worldwide in insulation, packaging and sponges — it randomly scatters into bubbles of varying size.

Wiebke Drenckhan, a CNRS researcher at the University of Paris South, and her colleagues now report a way to create plastics filled with ordered and nearly uniform bubbles. The researchers combine chemical reagents, surfactants, air and water in such a way that bubbles form and pack together in the liquid phase just before the surrounding material 'freezes' in a polymerization reaction.

Working with German chemicals company BASF, in Ludwigshafen, they have created bubble-stuffed foam sheets and threads that absorb water and can even be woven or knitted into fabrics (pictured below). Such foams might be used as membranes, acoustic filters or shear-resistant wraps for fibres containing carbon nanotubes.

ELSEVIER

THE ROYAL SOCIETY

Arboreal ascent

Proc. R. Soc. B doi:10.1098/rspb.2009.0911 (2009)

Vertebrates have been out on a limb for longer than previously thought, say Jörg Fröbisch at the Field Museum in Chicago, Illinois, and Robert Reisz of the University of Toronto at Mississauga in Canada.

Their preliminary description of the anatomy of *Suminia getmanovi*, a 260-million-year-old distant relative of mammals, concludes that the species represents the oldest evidence of a tree-dwelling vertebrate.

Comparison of the features of *S. getmanovi* fossils with those of modern reptiles and mammals reveals numerous features indicative of a life spent in trees. These include elongated limbs, a long and perhaps prehensile tail, and digits seemingly adapted for grasping and climbing, possibly with opposable thumbs.

NEUROSCIENCE

Learning experience

Neuron **63,** 244–253 (2009)

Sustained firing by neurons in two brain regions may help animals learn from the consequences of earlier actions.

Mark Histed, now at Harvard Medical School, and his colleagues trained macaques to move their eyes left or right in response to visual cues. The researchers located neurons in the prefrontal cortex and the caudate nucleus of the basal ganglia — two regions known to be involved in learning — that fired for several seconds after the monkeys found out whether their eye movements were correct.

The firing lasted until the following trial, suggesting that the neurons carried the link between the monkey's behaviour and its outcome into the next trial to facilitate learning.

GENETICS

Context is king

Science doi:10.1126/science.1174148 (2009)

In recent years geneticists have started looking at how genetic differences between individuals affect gene expression. Different levels of expression generally correlate with variations in regulatory genes.

© 2009 Macmillan Publishers Limited. All rights reserved

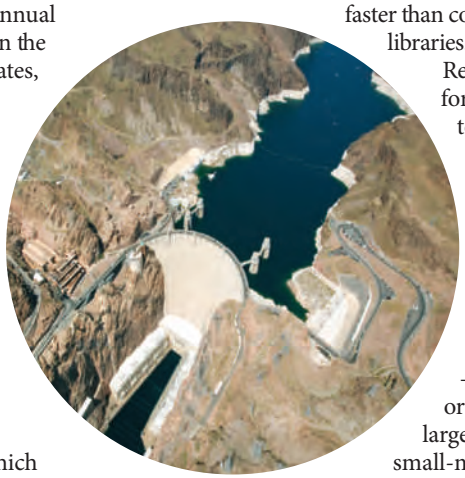
Emmanouil Dermitzakis and Stylianos Antonarakis, now both at the University of Geneva Medical School in Switzerland, and their team broke this down further by looking at how gene expression in different cell types derived from 75 people correlated with variations in their genomes. They took umbilical cord blood from pregnant women and cultured three types of cell for each. Comparing expression between individuals for each cell type, they found that 69–80% of gene variants affect expression levels in a manner specific to the cell type, suggesting that looking at just one tissue type is insufficient when comparing individuals.

WATER MANAGEMENT

Colorado be dammed

Water Resources Res. doi:10.1029/2008WR007652 (2009)

By 2057, dimate change could cause a tenfold increase in the annual risk of water shortages in the southwestern United States, say Balaji Rajagopalan at the University of Colorado in Boulder and his colleagues. Rajagopalan's team modelled a variety of management and climate scenarios for varying levels of potential demand over the period 2008–57 for the dam-created Lakes Powell and Mead (the latter pictured right), which together store Colorado River water for states including California, Arizona and Nevada.



THINKSTOCK/CORBIS

Most projections suggest that the river's flow will fall by 6–20% by the middle of the century. A 10% drop would mean a 25% chance of reservoirs being fully depleted on an annual basis, and a 20% drop would result in a 50% risk. Drought risk rises steeply after 2026, but management interventions could do much to reduce the risk of reservoir depletion, say the researchers.

INVERTEBRATE IMMUNITY

Infection in real time

PLoS Pathog. **5,** e1000518 (2009)

Scientists have a hard time following the initial action in an infection, but Will Wood at the University of Bath, UK, and his colleagues have tracked the early interactions between bacteria and immune cells as they battle for dominance in *Drosophila* embryos. They found that non-pathogenic *Escherichia coli* are successfully ingested by haemocytes, phagocytic immune cells that travel throughout the developing fruitfly.

However, the pathogenic bacterium *Photorhabdus asymbiotica* causes the haemocytes to freeze in place. By inserting or deleting specific host and pathogen genes, the authors showed that the haemocytes froze only when they engulfed a bacterial toxin called Mcf1, which seemed to have an effect on the immune cell cytoskeleton.

CHEMICAL BIOLOGY

800 million strong

Nature Chem. Biol. doi:10.1038/nchembio.211 (2009)

By tagging small molecules with short, double-stranded DNA fragments, Barry Morgan and his colleagues at GlaxoSmithKline in Waltham, Massachusetts, have created a collection of 800 million compounds that can be screened much faster than conventional chemical libraries.

Researchers can screen for ligands that bind to a protein target and then identify the new molecule by sequencing the DNA 'barcode' attached. As a proof of concept, the authors probed their DNA-tagged library — which is at least two orders of magnitude larger than a typical small-molecule library — for enzyme-blocking drug leads and pulled out novel inhibitors of two kinases.

STRUCTURAL BIOLOGY

Get into the groove

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0906532106 (2009).

A class of small molecules can distort the structure of DNA, disrupting its interaction with proteins that control gene expression.

Pyrrole–imidazole polyamides bind to the minor groove of the DNA double helix. David Chenoweth and Peter Dervan of the California Institute of Technology in Pasadena found that a cyclic polyamide binding in the minor groove compresses the major groove, which is where proteins called transcription factors bind DNA to regulate gene expression.

The polyamide also caused the helix to bend more than 18° towards the major groove, another distortion that could interfere with transcription-factor binding.

JOURNAL CLUB

Douglas Kell
The University of Manchester, UK

A systems biologist ponders how disparate ideas can sometimes come together beautifully.

If X alone and Y alone cannot explain a phenomenon, sometimes together they can. As the late biochemist Henrik Kacser remarked: "To understand the whole you must look at the whole."

Prion diseases, for example, are closely associated with the conformational change of the prion protein PrP from its normal form to an aggregating, autocatalysing, pathologic form, PrP^{Sc}. But clumping prions don't tell the whole story. Their levels often correlate poorly with disease progression, and it is far from clear how a simple conformational change leads to the holes in brain tissue seen in late-stage disease.

It is also clear that poorly liganded iron is highly neurotoxic, mainly because it can spur the production of the highly reactive and toxic hydroxyl radical $OH^{\bullet}$ — heavily involved in the progression of many other degenerative diseases and ageing. Neena Singh at Case Western Reserve University in Cleveland, Ohio, and her colleagues have now tied these two disparate threads together.

PrP^{Sc}, they found, can sequester cellular iron in insoluble PrP^{Sc}–ferritin complexes, making it bio-unavailable, leading to increased iron uptake and an overall excess of iron in brain tissue (A. Singh *et al.*, *PLoS Pathog.* **5,** e1000336; 2009). Modified iron metabolism is found in both scrapie and sporadic Creutzfeldt–Jakob disease, and such work stresses that it is not only the total amount of Fe(II) and Fe(III) that matters but their speciation. It is yet to be shown whether PrP^{Sc}–ferritin complexes catalyse $OH^{\bullet}$ production directly, but if they do, this could account for the massive damage observed. Recognition of this could have a colossal effect on our thinking and provide new therapeutic (and dietary) options based on iron chelation for these and other syndromes.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

© 2009 Macmillan Publishers Limited. All rights reserved

NEWS

US joins China in climate talks

But the two-day meeting was long on mutual understanding while being notably short on targets.

The United States and China adjourned a new round of bilateral talks in Washington DC last week with the vague outline of a climate partnership. But the 'G2' is far from sealing a meaningful deal in time for the global-warming summit in Copenhagen this December.

After two days of high-level talks opened by US President Barack Obama, officials from both countries signed a memorandum of understanding calling for cooperation on a range of energy and environmental issues. The document also establishes a bilateral dialogue covering the gamut of issues in the run-up to the United Nations climate talks.

Although much of the US–China meeting focused on monetary policy and security issues, energy and global warming emerged from the outset as a central theme in the talks, says Ken Lieberthal, a China expert based at the Brookings Institution in Washington.

Aside from a general call to "strengthen and coordinate" respective efforts on global warming, clean energy and other environmental initiatives, the memorandum itself was short on detail. It included no reference to programmes, targets or timelines for reducing greenhouse-gas emissions. The omission served as a tacit reminder that progress in the international climate talks will in no small part be determined by whatever comes of the bilateral negotiations between the world's top two polluters, together responsible for some 40% of global emissions.

US officials say they weren't expecting any breakthroughs from the meeting, which was designed to build relationships and deepen the discussion. "We're slogging ahead," says Todd Stern, the chief US climate negotiator. "But I do think that we will get there, and I think that there is a lot of interest on the Chinese side to arrive at a constructive and successful outcome in Copenhagen."

Chinese state councillor Dai Bingguo says that both countries face severe challenges and can make "an important contribution" to the global effort to tackle climate change. "We hope that through our joint efforts, we will be able to expand common ground and cooperation and take our collaborative efforts in these areas to a new height," Dai says.

The Strategic and Economic Dialogue began in 2006, but Obama and Chinese premier Hu Jintao have put the forum at the centre of their efforts to strengthen ties between the two increasingly interdependent countries. China flew about 150 diplomats into Washington this year, with the talks taking place almost exclusively behind closed doors.

M. WILSON/GETTY IMAGES

Warming measures: Chinese vice-premier Wang Qishan (left) and US president Barack Obama.

David Victor, a climate and international-policy expert at the University of California, San Diego, says the talks had a "false ring" under President George W. Bush, who withdrew the United States from the Kyoto Protocol on climate change. Victor acknowledges that the "ratio of talking to action" between the United States and China is still high, but says he believes the current negotiations are much more serious.

"The Chinese know absolutely that they have got to do something on climate," he says, suggesting that the Obama administration is worried about the domestic climate agenda as much as the international one. "The administration is scared that if they don't have anything credible with the Chinese and the Indians, that they are not going to be able to hold the politics together in the Senate."

The House of Representatives passed a comprehensive climate bill in June that included a cap-and-trade system to regulate greenhouse gases, but the outlook for passing something similar in the Senate remains unclear. Opponents argue that the United States can't solve the climate problem on its own, and thus should not bind itself to greenhouse-gas regulations unless major developing countries do so as well.

Nevertheless, William Chandler, a climate expert at the Carnegie Endowment for International Peace in Washington who has been working with Chinese officials, says the Chinese feel the international pressure and are committed to taking action on global warming. "China understands and accepts, I believe, that in the long run it will have to accept binding targets and caps on greenhouse-gas emissions," he says. "The important question is, 'What year will that be, and what do you do between now and then?'"

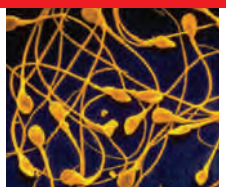
Chandler believes that Xie Zhenhua, who heads climate policy as vice-chairman of the National Development and Reform Commission, sees progress on global warming as a kind of personal legacy issue. Similarly, vice-premier Wang Qishan has taken a strong interest in global warming policy, according to Lieberthal. Both attended the talks in Washington.

"These happen to be two very talented individuals," Lieberthal says. "And they are both very committed to this issue." ■

Jeff Tollefson

See www.nature.com/roadtocopenhagen for more climate coverage.

© 2009 Macmillan Publishers Limited. All rights reserved



J. BURNS/DR RYDER/PHOTOTAKE

SPERM SPAT
Plagiarism accusation hits stem-cell research.
www.nature.com/news

Greek scientists fight research shake-up

Instead of enjoying a tranquil summer break, Greek researchers are fighting a major reorganization that will carve up two of the country's largest research centres.

Filippos Tsalidis, head of the development ministry's office for research and technology, took the scientific community by surprise by announcing the changes, intended to promote efficiency, in the business newspaper *Naftemporiki* on 3 June.

He plans to reshape research institutes overseen by his ministry, turning some into single-discipline centres and uniting smaller institutes in the south and north of the country into two regional centres.

Two major multidisciplinary research centres in Athens — the National Centre for Scientific Research Demokritos and the National Hellenic Research Foundation (NHRF) — will be partly dismantled. In 2001 and 2005, panels of international experts commissioned by the Greek government judged research at these centres to be poor, although improving in parts.

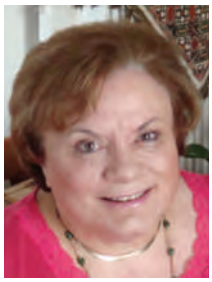
Greek scientists, angry at not being consulted about the restructuring, say it will cost more than it saves, is at odds with current multidisciplinary scientific trends, and will not solve the problem of underperforming research units. The plans have sparked public demonstrations, petitions and newspaper campaigns.

The opposition socialist party has pledged to reverse the reorganization if, as opinion polls predict, it wins power in the next election in 2010.

Under Tsalidis's plan, biology and organic chemistry institutes would transfer to a new facility that would be built at the elite Alexander Fleming Biomedical Research Centre outside Athens. Demokritos would absorb the NHRF's physics and remaining chemistry institutes, leaving the NHRF focused entirely on the humanities.

"It would be better to close institutes with poor evaluations rather than move them at great cost to a top-performing institute like the Fleming and dilute its efforts," says George Thireos, head of systems biology at the new Bioacademy research centre in Athens, which is run by the Academy of Athens and is not affected by the reorganization.

In 2006, Greece spent just 0.57% of its gross domestic product on research, one of the lowest percentages in the European Union, and there have been no competitive grants awarded for more than five years. "Changes to reduce duplication and promote collaboration are definitely needed here, but not in this way," says Effie Tsilibari, head of the Demokritos Institute of Biology, which faces relocation. "The proposal was rash, and it has left people paralysed."



"The proposal was rash, and it has left people paralysed."
— Effie Tsilibari

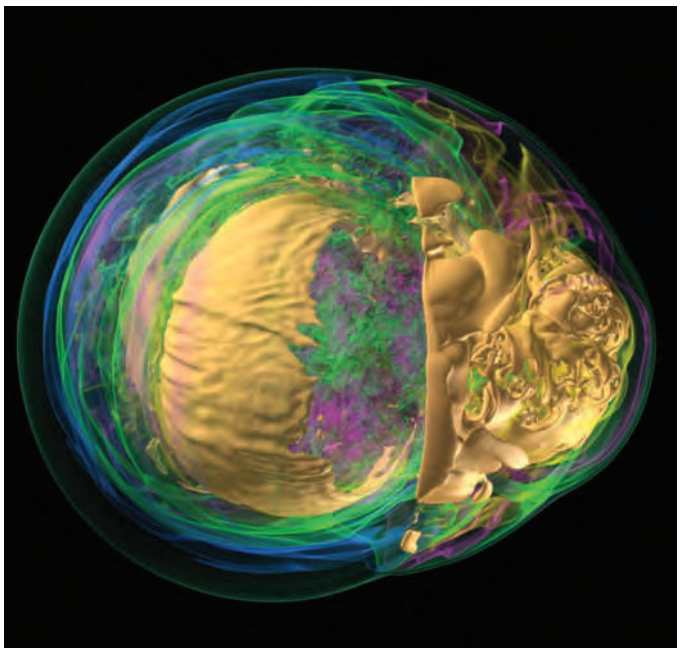
But George Kollias, director of the Fleming Centre, says the bombshell should be seen as an opportunity to stop talking and finally take action. "We should act even on this plan, using it to open a real discussion on how things that need to be changed can be changed," he says.

Senior Bioacademy scientists issued a statement on 30 July saying that although changes are needed, plans should be formulated through thoughtful discussions aimed at obtaining maximum consensus — and, above all, there must be greater financial investment. "Any plan that emerges without serious commitment of state funds will be doomed to failure," they warn.

Tsalidis did not respond to *Nature*'s request for an interview. ■

Alison Abbott

H. YU



SNAPSHOT

The guts of a dying star

This visualization of a simulated supernova is helping to reveal why pulsars spin so fast. Pulsars are neutron stars that emit beams of electromagnetic radiation as they whirl around many times per second — a rate that astrophysicists have struggled to explain. Pulsars are thought to form in core-collapse supernovae, the explosive deaths of stars at least eight times the mass of the Sun, in which the iron core collapses in on itself.

In 2007, computer simulations suggested that the stars don't explode in perfectly smooth spheres (J. M. Blondin and A. Mezzacappa *Nature* 445, 58–60; 2007). This latest visualization, created by Hongfeng Yu, a computer scientist at Sandia National Laboratories in Livermore, California, shows the entropy of the gases in the dying star's core, revealing the immense swirling currents that originated as tiny perturbations (gases with the highest entropy are yellow, followed by green and then purple). The currents "spin up the proto-neutron star, just like pulling a string on an old spinning top", says Bronson Messer, an astrophysicist at Oak Ridge National Laboratory in Tennessee, who contributed to the research. The work incorporates a new visualization technique, developed at Argonne National Laboratory outside Chicago, Illinois, which runs and visualizes the simulation directly on a Blue Gene/P supercomputer.

Lizzie Buchen

© 2009 Macmillan Publishers Limited. All rights reserved

SPECIAL REPORT

Who speaks for science in Europe?

Questions remain over whether researchers have a coherent enough voice to influence European science policy. **Natasha Gilbert** reports.

European scientists don't have too much to complain about these days. More than €50 billion (US$70 billion) is budgeted for research in the European Union's (EU) Seventh Framework Programme (FP7), which began in 2007 and runs until 2013. And two years ago, scientists saw the realization of their long-sought dream of a European Research Council (ERC) that would fund frontier research judged solely on excellence.

But still researchers talk about how things could be better. As the debate begins on what the next framework programme should look like (see 'Getting input'), they are asking how they can get their voices heard more effectively.

The problem is how to work effectively within Europe's notorious but necessary constitutional bureaucracy. The European Commission draws up proposals for the framework programme, including the areas of research that it covers, how much money it gets and initiatives such as the ERC. But member states and the European Parliament must give the commission's proposals the green light.

The commission thus has to accommodate a wide range of competing views, including differing national and political priorities as well as the wishes of scientists and industry, says Achilleas Mitsos, who headed the commission's research directorate-general, based in Brussels, between 2000 and 2005. Amid the din, the input of scientists can get lost if it is not provided coherently through an influential outlet.

"There is no homogeneous body through which scientists can speak with one voice," says Dieter Imboden, president of EUROHORCs, a group of the heads of European research councils.

European champion

Some new ideas may help. One that is beginning to gain ground is the possibility of having a chief scientific adviser for Europe, who would inform policy on such matters as energy or genetically modified crops. Insiders say discussions on what such an adviser could do, and how the position could be structured, are expected to begin next month. In theory, this person could serve as a single focal point for scientific input. John Beddington and David King, the current and former science advisers to the UK government, have both publicly touted this idea.

Another idea is to revise the themed approach of the current framework programme to focus instead on broader societal questions. Discussions on this are expected to start next year. At a conference in Lund, Sweden, last month, around 350 researchers and politicians agreed on a joint statement, known as the Lund Declaration, that calls for European research funding to focus on "grand challenges" rather than the "rigid thematic" approach of FP7. The declaration says that academia must have a larger role in identifying these challenges.

Currently, scores of disparate groups, from university organizations and science bodies to research institutions and individual scientists, send suggestions to the commission to help it formulate ideas about what the framework programmes should include. Once the proposals are drawn up, scientists can also comment on them, to a degree that can be almost overwhelming. For example, for FP7 the commission published on its website the opinions of 34 stakeholder groups from academia and industry.

But Helga Nowotny, a vice-president of the ERC's scientific council, says it is not clear what happens to this advice. The way the framework programmes are put together, she says, is not particularly transparent.

F. LENOIR/REUTERS

GETTING INPUT

The framework programmes (FPs) that guide European research funding priorities are put together in a complex process involving multiple stakeholders — of whom scientists are but one.

2004 Discussion formally starts on FP7, which will distribute more than €50 billion for research

2005 The European Commission publishes broad outlines of FP7 and requests feedback

2006 The FP7 proposals are amended on the basis of feedback, then approved by the European parliament and European council

2007 FP7 begins

2010 Formal discussions on FP7's successor, FP8, are likely to start

2013 FP7 ends and FP8 begins

Group practice

When it comes to the planning of work for each year of a framework programme, another set of scientific advisers rolls into action. The annual work programmes set out details for specific research topics: for instance, FP7 contains a general commitment to cardiovascular research, but only in the work programme for 2009 is it specified that funding will go to investigating the genetic and environmental factors influencing cardiac arrhythmia.

Fourteen permanent science advisory groups help draw up these work programmes; each covers a different research area, including energy, the environment, and information and communication technologies (ICT). Most members are scientists or representatives of industry, all selected by the commission.

The remit and influence of the advisory groups vary considerably. Michael Depledge of the University of Plymouth, UK, chairman of the environmental and climate-change science advisory group, says he has been "frustrated" over what his 20-person group can do. When he joined in 2007 as vice-chairman, the

© 2009 Macmillan Publishers Limited. All rights reserved

CUTS TO ISRAEL'S SPACE INDUSTRY
Drop in funding triggers crisis.
www.nature.com/news

IAI

European Union bureaucracy dogs research policy.

commission had limited the group to commenting on draft work programmes that the commission had already drawn up. When the group wanted to make changes, such as altering the order in which calls for proposals were advertised, the commission told them it was impossible because the programme had already been agreed by member states.

"We ended up just going through the text of the work programme commenting on individual words that may need changing," says Depledge. "The whole group found this completely unsatisfactory. It was just a box-ticking exercise."

When Depledge became chairman six months ago, he decided to shake things up. The group abandoned picking through work programmes and began providing advice on emerging issues that should be covered in future Framework programmes. Depledge says the group will produce a report at the end of this year, and another next year, on the direction it thinks environmental research needs to go. "The commission will then be faced with a situation where, if they ignore our advice, they will have to explain why they did so," he says.

His experience contrasts with that of Chris Hankin, a computer scientist at Imperial College London and a member of the nearly-40-person advisory group for ICT. Unlike Depledge, whose group meets four times a year, Hankin says he is in Brussels nearly every month. He is currently involved in two projects for the group: one looking at what ICT research should be funded in the last two years of FP7, and the other on future and emerging technologies. The group has also produced reports on how ICT is likely to change over the next ten years. In all these cases, the work has been done because the commission requested advice. "We are involved much earlier in the process of developing the work programmes than other advisory groups," Hankin says.

This is because the groups deal with different sections of the commission, he says, some of which are "more proactive in seeking advice", and because his group has a long history of being active.

Historical inertia

Overall, however, inadequate input from scientists has left successive Framework programmes "dominated by historical inertia", says Luke Georghiou, an expert in European science policy at Manchester University, UK. "If a research area is already in, it will tend to stay in the next programme," he says. For instance, his analyses suggest that the proportion of funds for research allocated to each of the nine high-level priorities of the first six framework programmes have stayed essentially constant since the first one began in 1984.

Imboden says scientists need to take responsibility for the "deficiency" in their contributions. "I don't think there is much reason to blame the commission here," he says. A different opinion is voiced in an independent evaluation of FP6, which ran from 2002 to 2006. The report, published in February, concluded that "more transparent consultation with stakeholder communities … would have produced a more robust overall FP design".

"There is no homogeneous body through which scientists can speak with one voice."

Mitsos says the commission has always interacted with the broader scientific community, including holding meetings with various university and academic bodies. He cites the creation of the ERC as a good idea that sprang from the community. "I don't accept that the channels of communication are not there," he says. "The question is, to what extent is the commission listening or is able to transform what it hears into specific proposals?"

Janez Potočnik, the EU research commissioner, says the commission needs "more clear, strategic thinking about what is in front of us". How to do that will take more work, he says; he favours discussing the idea of a European chief science adviser, but notes that having one key person as contact could potentially shut out other scientific opinions. And as for revising the way framework programmes are put together, he says, "premature debate over FP8 would hinder rather than help".

For advice on longer-term science strategies, the commission has its own advisory body of 22 scientists and science-policy experts, now called the European Research Area Board (ERAB). But some question its effectiveness. Nowotny, who chaired the body in its former life as the European Research Advisory Board, says the commission didn't use the board much. "Rarely did the commission ask us for advice on specific issues," she says.

According to John Wood of Imperial College London, the chairman of ERAB, the commission is now readier to ask. For example, the commission recently asked ERAB to "think outside the box" on what the EU research landscape should look like. Its report is due to be published in September, and will include recommendations on the use of independent science advice, future Framework programmes and how funds for research should be distributed.

Wood says the board has a direct link to the commission because he regularly meets with Potočnik. But he is not sure what happens to its advice after this. "It's very opaque," he says.

In addition to its permanent groups that provide advice, the commission also sets up *ad hoc* advisory bodies to help it. Georghiou has been involved in a number of these and says their recommendations have been followed to varying extents. "Success factors are having the right message at the right time and making sure it is delivered," he says.

Having the ear of influential people is also crucial. Last year, for example, Georghiou chaired an expert group for the commission which suggested, as the Lund Declaration did, that funding should focus on grand challenges. "It seems likely that the next framework programme will look very different and will have this grand-challenge element," he says.

Yet the question still remains, Imboden says, of who speaks for science in Europe. One answer might be beginning to emerge: on 24 June, Georghiou co-chaired a meeting of a pilot forum that brought together the key EU science and academic bodies, including ERAB, EUROHORCs and the European University Association. The meeting discussed a vision for the European Research Area, including what the next FP should look like.

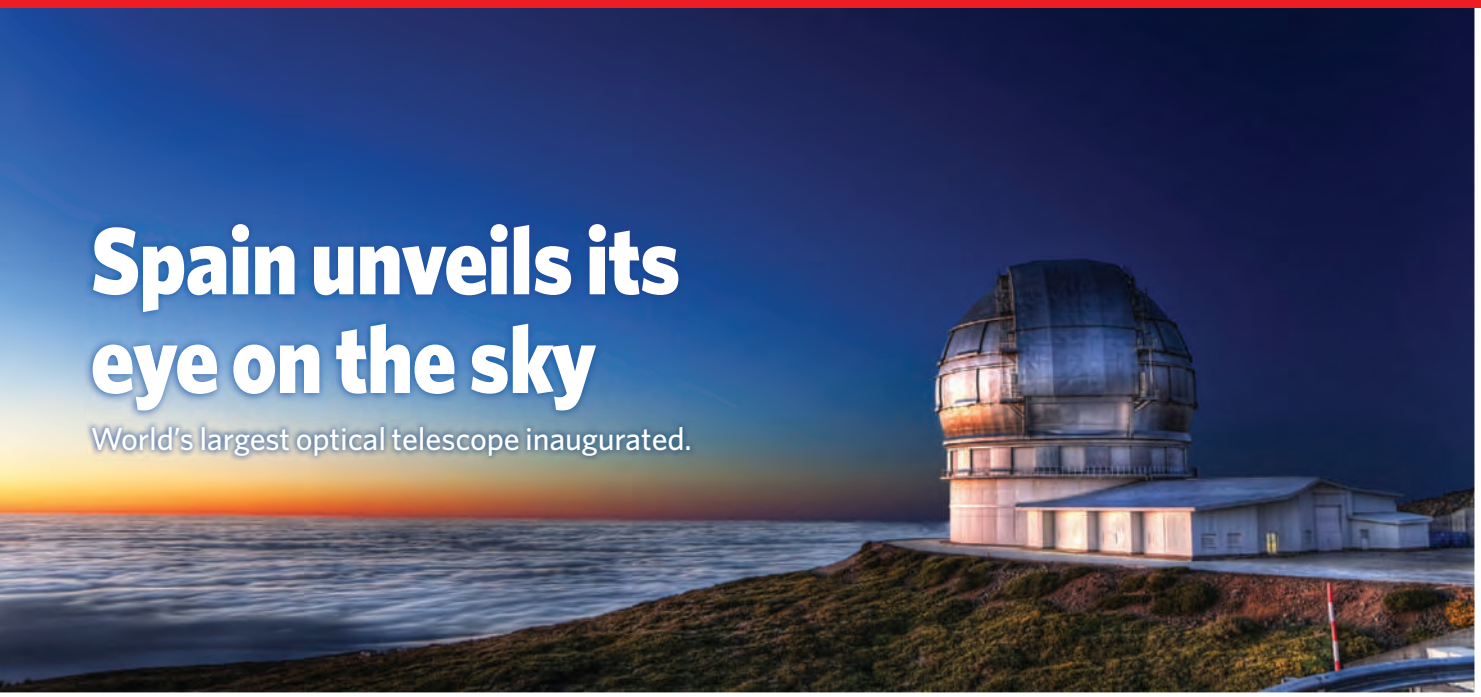
There is "strong support" for the forum to become a more permanent body, Georghiou says. And that could end up as the space for Europe's disparate science voices to come together as a coherent whole. ■

Natasha Gilbert

© 2009 Macmillan Publishers Limited. All rights reserved

Spain unveils its eye on the sky

World's largest optical telescope inaugurated.



P. BONET

LA PALMA, SPAIN

As the world's largest single optical telescope officially opens for business, some astronomers are still wondering precisely what that business should be.

The Gran Telescopio Canarias (GTC), which boasts a 10.4-metre mirror composed of 36 hexagonal segments, is the latest addition to the Roque de los Muchachos Observatory perched about 2,400 metres above sea level on La Palma, one of the Spanish Canary Islands.

The telescope's 24 July inauguration by King Juan Carlos of Spain attracted astronomers from around the globe, who stayed on for a two-day scientific symposium. Most agree that the telescope is a valuable addition to the 8- to 10-metre class of telescopes, such as the twin 10-metre Keck telescopes in Hawaii. Some argue that the GTC also makes an ideal testbed for the technology and instruments needed by its successors — monster telescopes with mirrors measuring 25, 30 or even 42 metres, which are scheduled to come online in the coming decade (see *Nature* **452,** 142–145; 2008).

"It will surpass the quality of the Keck telescopes, thanks to better stability and better optics."

But others have their doubts about how much the GTC can achieve before then. "Keck picked much of the low-hanging fruit," concedes William Smith, president of the Association of Universities for Research in Astronomy. Keck pioneered the search for remote galaxies, for example, before the GTC was conceived.

Bruno Leibundgut, director for science at the European Southern Observatory (ESO), which operates the Very Large Telescope in Chile, thinks the Spanish newcomer "needs to find a niche where it can provide something that other telescopes haven't done yet". He suggests that the GTC could undertake time-consuming survey projects or specialized follow-up observations of, say, γ-ray bursts spotted by orbiting telescopes.

But Rafael Rebolo, a research professor at the Canaries Institute of Astrophysics (IAC) in Tenerife, thinks that the GTC will be opening up new astronomical domains, studying the very first galaxies in the Universe or carrying out mid-infrared observations of cool extrasolar planets.

The GTC's observing programme will depend in part on what instruments are installed. An imaging spectrograph called OSIRIS is currently the GTC's only operational instrument; an infrared camera called CanariCam is still in boxes awaiting installation later this year. GTC director Pedro Álvarez says that they plan to install a diverse set of astronomical instruments over the next 3–4 years, including a near-infrared spectrograph that can study many objects simultaneously, a high-resolution spectrograph for visible-wavelength observations and a near-infrared camera that will use a planned adaptive optics system to mitigate atmospheric turbulence.

The GTC was originally supposed to become operational in 2003. Francisco Sánchez, director of the IAC, initiated the GTC project in 1998 and admits that their schedule was too optimistic.

The effort initially met with strong scepticism — Spain had little experience in building optical telescopes, its previous largest being just 80 centimetres wide (see *Nature* **435,** 140–142; 2005). "When a bicycle repair man announces he's going to build a Porsche, you're naturally unconvinced," says René Rutten, GTC head of astronomy operations. "However, I now believe it will surpass the quality of the Keck telescopes, thanks to better stability and better optics."

"My main worry is to quickly provide the best possible instrumentation," adds Sánchez. "Given the imminent emergence of extremely large telescopes, our window of opportunity is small." Indeed, the Thirty Meter Telescope — joining the Keck telescopes atop Mauna Kea in Hawaii — should be operational by 2018 (see *Nature* **460,** 563; 2009).

Timeshare

Because 90% of the €105-million (US$150-million) GTC budget was provided by Spain, most of the telescope's observing time will go to Spanish astronomers. But Rebolo says the GTC is still open to new partners, and other European astronomers will get access to the GTC thanks to Spain's membership of the ESO consortium. When Spain joined the ESO in 2006, it offset about a quarter of its €65-million entrance fee with the promise of a total of 122 observing nights on the GTC for astronomers from consortium member states. The deal also included 55 'technology days', allowing ESO engineers to get hands-on experience with the GTC's segmented mirror — potentially useful in designing and constructing the proposed 42-metre European Extremely Large Telescope (E-ELT).

According to Sánchez, this synergy could become even stronger if the E-ELT were built at the Roque de los Muchachos Observatory, which is one of a handful of candidate sites under consideration. "Building the E-ELT at La Palma would favour the further development of telescopes here, and would enormously promote European astronomy," agrees Rebolo.

E-ELT construction could also be speeded up by choosing La Palma as its location. Spain is more than willing to invest in the project, and the European Union could sponsor it with extra funds provided through its Ultra Peripheral Regions development programme. Moreover, of the six possible sites currently under review, La Palma is the only one with an existing infrastructure of roads and support buildings. "It's only natural to host the E-ELT here," says Sánchez. ■

Govert Schilling

© 2009 Macmillan Publishers Limited. All rights reserved

Joint Mars plans come together

NASA and the European Space Agency (ESA) have unveiled a joint plan for exploring Mars in the latter half of the next decade. ESA will build a trace-gas orbiter, able to map plumes of methane in the atmosphere, for launch in 2016. This could help to target landing of the agency's flagship rover, ExoMars, and a mid-sized NASA rover, due for launch in 2018.

"These two rovers will be focused on astrobiology — seeking the signs of life," says NASA's Mars programme chief Doug McCuistion, who told the US Mars community about the plan at an advisory committee meeting on 29 July at Brown University in Providence, Rhode Island.

The plan was negotiated at a NASA–ESA summit in Plymouth, UK, at the end of June, and McCuistion says that ESA member states have now agreed to it. Jack Mustard, a Brown University geologist and chair of a NASA Mars advisory group, says that the community is pleased to have a 2016 orbital mission at all after the Mars Science Laboratory (MSL), a large rover still scheduled for a 2011 launch, ran roughshod over NASA budgets with its price tag, which could end up being as high as $2.4 billion. "Scientists are definitely happy to have a viable opportunity for measurements," he says. "But it's far too new. No one knows what it means or how it's going to work out."

Each agency has negotiated its share of the work. NASA will provide Atlas rockets for both launches, and in 2018 it aims to re-use the 'sky crane' technology that it is developing to lower the MSL to the planet's surface. Initially, NASA and ESA officials hoped to squeeze ExoMars and a trace-gas orbiter onto the same Atlas rocket for a 2016 launch. But ESA eventually agreed to delay ExoMars until 2018, a launch window with better orbital mechanics.

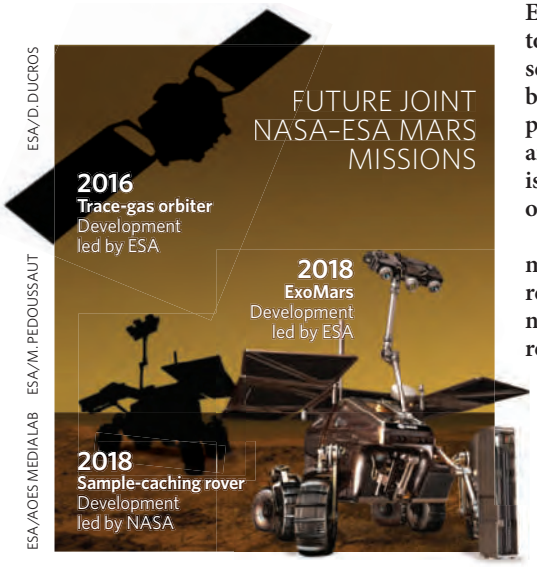
ESA plans to use leftover payload on the 2016 rocket for a small lander as a way to test tricky technology for entry, descent and landing. In 2018, NASA is looking to fit another rover on board. Whereas ExoMars will drill cores as much as two metres deep to look for life, the new NASA rover — bigger than the current rovers Spirit and Opportunity, but smaller than the MSL — would analyse and cache rocks as a first step in a far-off sample-return mission.

Both could be aided by the 2016 orbiter, if it were able to direct the rovers to a landing site near vents of methane, which can be produced by subterranean microbes or by hydrothermal processes on certain volcanic rocks. A paper in *Nature* this week (see page 720) shows that the observation of seasonal methane plumes cannot be explained by conventional models for atmospheric circulation, which should disperse the methane uniformly. The authors instead posit that seasonal plumes of methane can exist only if the gas is destroyed quickly in surface interactions with soils.

A joint NASA–ESA science team has just begun working out the orbiter's design requirements, with instruments expected to be awarded competitively to ESA or NASA scientists. Sushil Atreya, an atmospheric scientist at the University of Michigan in Ann Arbor who has worked on ESA's Mars Express orbital mission, says he is pleased to see initial designs calling for methane sensitivities of parts per trillion. That would be orders of magnitude better than the parts-per-billion measurements of Mars Express, and could allow the detection of light carbon isotopes — a possible indicator of biological origin — within methane molecules.

But mapping the methane will be much, much tougher, says Atreya. A spatial resolution of 10 or 20 kilometres would be necessary to be of any help in targeting a rover. Mars Express was only able to find hints of regional methane variability, whereas the plumes discovered by ground-based observations were discerned across hundreds of kilometres. "Have you ever tried to catch gas in the wind?" asks Mustard. "It's a moving target." ■

Eric Hand



ESA/D.DUCROS ESA/M.PEDOUSSAUT ESA/AOES MEDIALAB

© 2009 Macmillan Publishers Limited. All rights reserved

Grant scores leave applicants in limbo

Applicants for the coveted Challenge Grants issued by the US National Institutes of Health (NIH) as part of the American Recovery and Reinvestment Act learned the peer-review scores for their proposals late last month. Yet they received little in the way of certainty over whether those scores will translate into money come September, when the NIH will announce which grants it plans to fund.

Competition for the US$1-million, two-year awards is fierce — the agency in Bethesda, Maryland, received more than 21,000 applications, and the NIH director's office will fund only about 1% of these.

"I don't think I've ever been ambivalent about a second percentile score" that would normally be assured funding, says Joe Hogan, a biostatistician at Brown University in Providence, Rhode Island, who hopes to use a Challenge Grant to study behavioural interventions for reducing alcohol abuse.

"I don't think I've ever been ambivalent about a second percentile score."

With ordinary grants, applicants can usually tell if their grant is fundable as soon they receive their percentile score because they already know the designated 'payline', or percentage of fundable applications. The NIH has designated an initial $200 million of $10.4 billion in economic stimulus funds for the grants, but with so many variables at play in allocating the stimulus money, predicting whether a given score will land funding is almost impossible — meaning that those with percentile scores in the mid-single digits are left hanging.

For example, Paul Janssen, a cardiac physiologist at Ohio State University in Columbus, scored a sixth percentile on his 'infrastructure' application, which, if funded, would build a system for obtaining and testing live tissue from healthy and failing human hearts. "I am cautiously optimistic," he says, only because the institute sponsoring the grant — the National Heart, Lung and Blood Institute — is planning to go beyond the allocation from the director's office and fund 200 Challenge Grants in its topic areas. Some of the 27 institutes within the NIH are less enthusiastic about funding extra Challenge Grants and have chosen to use stimulus funds in other ways — for example, to boost existing investigator-initiated grants, or to sponsor standard grants that had fallen just short of the payline before the stimulus windfall arrived.

Meanwhile, the burden on the thousands of grant reviewers has, according to some, turned out to be bearable. Gary Johnson, chairman of the pharmacology department at the University of North Carolina, Chapel Hill, told the NIH that he could review up to five Challenge Grant applications. "And they only gave me a couple," he says. "I don't know anyone who was overwhelmed by reviewing, because there was an overwhelming agreement of investigators to participate in the process." ■

Meredith Wadman

© 2009 Macmillan Publishers Limited. All rights reserved

India embarks on push to become a solar power

India's prime minister Manmohan Singh has unveiled a 30-year, US$19-billion plan to make the country a leader in solar energy.

Announced on 3 August, the programme aims to raise installed solar capacity from its current 5 MW to 20 GW by 2020, 100 GW by 2030 and 200 GW by 2050, although a detailed road map has been drawn up to 2020 only. An autonomous solar-energy authority will be created to execute the mission, and the existing solar-energy centre near New Delhi will be upgraded to an institute that will coordinate solar-research centres across the country and promote foreign collaboration — a key feature of the plan.

Industry carrots include tax credits and priority bank loans for solar-power projects, as well as the duty-free import of raw materials. And conventional power plants with steam-driven turbines will have to generate at least 5% of their capacity from solar power.

For a longer version of this story, see http://tiny.cc/iem9l

Lab worker charged with destroying protein crystals

A former employee who allegedly destroyed US$500,000 worth of protein crystal samples at the SLAC National Accelerator Laboratory in Menlo Park, California, was arrested and charged last week with wilfully ruining government property.

Silvya Oommachen, until July a research associate at SLAC's Joint Center for Structural Genomics (JCSG), removed 4,000–5,000 protein crystals from three SLAC freezers at some point between 17 and 20 July, according to an FBI affidavit.

The now-useless crystals formed part of the Protein Structure Initiative, a federally funded project to expedite the discovery of atomic-level protein structures. JCSG director Ian Wilson estimates that his research team now faces a "two- to three-month setback" to remake the protein crystals that had not yet been analysed.

For a longer version of this story, see http://tinyurl.com/lfnj43

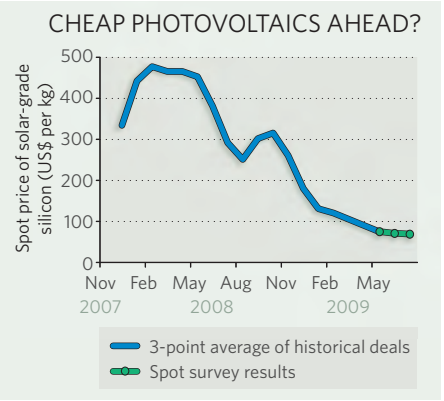
Plummeting silicon prices may boost solar sales

The price of silicon for the solar-power industry has plunged in the past year as a result of increasing supplies and a sharp drop in demand, with the price of silicon photovoltaic panels poised to follow.

The spot price of solar-grade silicon — for immediate delivery — has fallen by roughly 77%, from an average of more than US$300 per kilogram last year to $67 per kilogram today, according to the London-based consultancy New Energy Finance (see graph).

That has forced silicon and wafer suppliers to renegotiate contracts signed last year for delivery this year. Contracts signed at $150 per kilogram have been cancelled or renegotiated at roughly 50% discount.

Even before the global financial crisis, analysts had warned that supplies would outstrip demand in 2009, with new manufacturing facilities coming online at a time when countries such as Spain are scaling back solar subsidies. The good news for solar manufacturers, the consultancy reports, is that they should be able to halve the price of panels, which should spur demand.



SOURCE: NEW ENERGY FINANCE

US report backs distinction between science and policy

In setting regulatory policy, the US government should do more to separate scientific advice from policy decisions based on that advice, according to a report released on 5 August by the Bipartisan Policy Center, a non-profit body based in Washington DC established by former Democratic and Republican members of Congress.

The report recommends that regulatory agencies should post public notices that distinguish between the science and policy questions being asked. In appointing scientific advisory panels, agencies should adopt more stringent requirements about financial and professional conflicts of interest, and should be more transparent in disclosing them, it says. The report has been reviewed by the administration of President Barack Obama, who condemned the politicization of science in March, and is expected to issue guidelines on these issues later this year.

Chikyu showcases riser drilling for deep-sea research

IODP/JAMSTEC

The first scientific ocean-floor drilling project to use a riser drill — equipment previously used in oil exploration — was completed last week.

The Japanese research vessel *Chikyu* (pictured) drilled 1,600 metres below the sea floor of the Nankai Trough, an earthquake-generating zone off the Pacific coast of Japan.

Riser drilling circulates mud in an extra casing around the drill to prevent the collapse of a borehole in deep, high-pressure zones. *Chikyu* had already tested its riser-drilling equipment while on loan to an Australian oil company (see *Nature* **442,** 964; 2006).

The vessel is taking a leading role in the Integrated Ocean Drilling Program, a collaboration of Japanese, US and European scientists studying rock and sediment samples to learn about Earth's structure and history. It is due to drill two more sites in the Nankai Trough.

Corrections

The News story 'Biodefence lab criticized' (*Nature* **460,** 556–557; 2009) conflated two different foot-and-mouth disease outbreaks in Britain. The 2001 outbreak required the slaughter of 6 million animals; a 2007 outbreak originated from the animal-research lab in Pirbright.

The News story 'Flu jabs urged for developing countries' (*Nature* **460,** 156–157; 2009) incorrectly stated that Abdullah Brooks has determined that one-third of pneumonia deaths in children younger than 2 years old in Bangladesh can be attributed to the influenza virus. In fact, he has determined that about one-third of children who get influenza develop pneumonia, of whom about two-thirds are less than two years old.

© 2009 Macmillan Publishers Limited. All rights reserved

Sending out an SOS

An Obama gambit on space policy highlights the benefits and risks of turning to outside experts. **David Goldston** explains.

PARTY OF ONE

When a public official appoints a commission to study a problem, it's usually assumed to be a craven strategy to delay a decision or to dodge responsibility. But sometimes tossing an issue over to outsiders actually offers the best chance of moving forward. That seems to be the case with the commission appointed in June by President Barack Obama's administration to review NASA's human spaceflight programme. Without such a panel, which is due to report late this month, developing a workable consensus on NASA's future would probably be impossible. The only problem is that reaching consensus might still be impossible even with the panel.

The panel, chaired by Norman Augustine, the widely respected former head of aerospace giant Lockheed Martin, is charged with figuring out how to put together a viable human spaceflight programme within the existing budget. No one thinks the United States has such a programme now. Both the budget and schedule for President George W. Bush's initiative for a US return to the Moon seem impossibly tight, and its spacecraft and rocket designs have always been controversial. Moreover, congressional supporters of the programme have never truly accepted the decision to cancel the space shuttle in 2010 and to abandon the International Space Station in 2016 to free up funds for the lunar project.

It didn't take Obama administration officials long to recognize that the status quo was unsustainable. Worse still, they realized, the budgetary and political instability in NASA's human spaceflight programme threatened their express commitments to rebuild the agency's Earth-science capabilities and to put its space-science programme, plagued by its own cost overruns, on a sound footing. But the administration had no clear path forward and little time. Shortly after taking office in January, it had to make decisions on budgets for fiscal years 2009 and 2010, not to mention an economic stimulus package.

The key officials handling the problem — presidential science adviser John Holdren, acting NASA administrator Christopher Scolese, their staffs and the staff at the Office of Management and Budget — began to see a commission not only as the best way to move ahead, but as the sole way. They realized that the Obama administration on its own would never have the credibility to forge a new consensus with NASA's supporters. The administration would be too vulnerable to a range of disparate charges: that it was a captive of NASA, or not genuinely committed to space exploration, or merely interested in reversing Bush policy, or just trying to make the budget arithmetic work regardless of the consequences.

The administration did genuinely need the technical guidance a commission would provide — NASA couldn't be expected to objectively review its own work — but the primary motivation for appointing the Augustine panel was, legitimately, political. A respected group of experts with clout on Capitol Hill might be able to reshape the debate over NASA's future and tamp down the carping and second-guessing that had afflicted the human spaceflight programme and made its budget a perpetually open question.

A commission was good strategy within the White House as well. As a presidential candidate, Obama had eventually endorsed the lunar mission, but like most presidential hopefuls, space had hardly been his primary focus or interest. (Even President John F. Kennedy reportedly admitted to NASA officials, "I'm not that interested in space".) Yet Holdren and his colleagues realized that the nature of the space programme would have to be a presidential decision; fundamental policy choices with political and strategic consequences were involved. The Augustine panel, with its high-powered members and very public operation, would help get the president's attention. It would also give him more information on which to make a decision and confidence that all the options and angles had been reviewed. A commission was needed to gain credibility within the administration, not just outside it.

The only downside of such an approach was both basic and unavoidable: it wasn't clear what the panel would conclude. Turning to the commission, officials acknowledge, was something of a gamble, a 'crap shoot'.

If all goes well from the administration's point of view, the Augustine committee will design an exploration programme that can satisfy space enthusiasts while not veering far from its official charge to stay within the existing budget. Already everyone expects the group to offer several scenarios, with a number of them exceeding current spending targets. Panel members have talked seriously about extending both the space shuttle and station programmes, which each carry hefty price tags, and many experts believe the lunar programme needs significantly more cash to succeed.

If the higher-cost options are the only ones that win favour with human spaceflight advocates, the administration may be worse off than it would have been without the Augustine panel. Supporters of increased spending will have been given fresh ammunition and the enhanced legitimacy that the panel was supposed to provide the administration. If the costs are a lot higher, the president will be forced to reassess his commitment to returning to the Moon, his commitment to strengthen NASA's science programmes, or his effort to constrain the agency budget in a period of burgeoning deficits. The ploy will have backfired, opening a Pandora's box of fundamental questions the White House was quite reasonably trying to avoid.

The outcome of such a re-examination and the ensuing debate is difficult to predict. Space-policy issues have rarely been front-page news over the past 40 years, and even Bush's decision to return to the Moon faded quickly from public view and failed to generate much congressional discussion.

But a revised space policy now, at a time of high unemployment, soaring budget deficits and pent-up demands for scientific research, might lead to a more fully engaged debate about what NASA's priorities should be. The most likely result would be yet another armistice between NASA's human spaceflight and science programmes, leaving both insufficiently funded. For now, though, a genuine, broad and open conversation about NASA's focus remains space policy's final frontier. ■

David Goldston (partyofonecolumn@gmail.com) is the director of government affairs at the Natural Resources Defense Council in Washington DC. Views expressed are his own.

© 2009 Macmillan Publishers Limited. All rights reserved

NATURE|Vol 460|6 August 2009

Meltdown modelling

Could agent-based computer models prevent another financial crisis? **Mark Buchanan** reports.

It's 2016, and experts at a US government facility have detected a threat to national security. A screen on the wall maps the world's largest financial players — banks, governments and hedge funds — as well as the web of loans, ownership stakes and other legal claims that link them. High-powered computers have been using these enormous volumes of data to run through scenarios that flush out unexpected risks. And this morning they have triggered an alarm.

RECESSION WATCH

Flashing orange alerts on the screen show that a cluster of US-based hedge funds has unknowingly taken large ownership positions in similar assets. If one of the funds should have to sell assets to raise cash, the computers warn, its action could drive down the assets' value and force others to start selling their own holdings in a self-amplifying downward spiral. Many of the funds could be bankrupt within 30 minutes, creating a threat to the entire financial system. Armed with this information, financial authorities step in to orchestrate a controlled elimination of the dangerous tangle.

Alas, this story is likely to remain fiction. No government was able to carry out any such 'war room' analyses as the current financial crisis emerged, nor does the capability exist today. Yet a growing number of scientists insist that something like it is needed if society is to avoid similar crises in future.

Financial regulators do not have the tools they need to predict and prevent meltdowns, says physicist-turned-sociologist Dirk Helbing of the Swiss Federal Institute of Technology Zurich, who has spent the past two decades modelling large-scale human systems such as urban traffic or pedestrian flows. They can do a good job of tracking an economy using the statistical measures of standard econometrics, as long as the influences on the economy are independent of each other, and the past remains a reliable guide to the future. But the recent financial collapse was a 'systemic' meltdown, in which intertwined breakdowns in housing, banking and many other sectors conspired to destabilize the system as a whole. And the past has been anything but a reliable guide of late: witness how US analysts were led astray by decades of data suggesting that housing values would never simultaneously fall across the nation.

> **"We have had a massive failure of the dominant economic model."**
> — Eric Weinstein

Likewise, economists can get reasonably good insights by assuming that human behaviour leads to stable, self-regulating markets, with the prices of stocks, houses and other things never departing too far from equilibrium. But 'stability' is a word few would use to describe the chaotic markets of the past few years, when complex, nonlinear feedbacks fuelled the boom and bust of the dot-com and housing bubbles, and when banks took extreme risks in pursuit of ever higher profits.

In an effort to deal with such messy realities, a few economists — often working with physicists and others outside the economic mainstream — have spent the past decade or so exploring 'agent-based' models that make only minimal assumptions about human behaviour or inherent market stability (see page 685). The idea is to build a virtual market in a computer and populate it with artificially intelligent bits of software — 'agents' — that interact with one another much as people do in a real market. The computer then lets the overall behaviour of the market emerge from the actions of the individual agents, without presupposing the result.

Agent-based models have roots dating back to the 1940s and the first 'cellular automata', which were essentially just simulated grids of on–off switches that interacted with their nearest neighbours. But they didn't spark much interest beyond the physical-science community until the 1990s, when advances in computer power began to make realistic social simulations more feasible. Since then they have found increasing use in problems such as traffic flow and the spread of infectious diseases (see page 687). Indeed, points out Helbing, agent-based models are the social-science analogue of the computational simulations now routinely used elsewhere in science to explore complex nonlinear processes such as the global climate.

ILLUSTRATIONS BY JESSE LEFKOWITZ

© 2009 Macmillan Publishers Limited. All rights reserved

That is why he is eager to bring social and physical scientists together to develop computational 'wind tunnels' that would allow regulators to test policies before putting them into practice. "The idea is to invest a lot in science," he says, "and thereby save hundreds of times as much by avoiding or mitigating future crises."

Just more theory?

That notion is a tough sell among mainstream economists, many of whom are less than thrilled by offers of outside help. "After any crisis," says Paul Romer of Stanford University, California, a leading researcher in the economics of innovation, "you hear recommendations to recruit scientists from other fields who can purge economics and finance of ideology and failed assumptions. But we should ask if there is any evidence that more theory, developed by people who don't have domain experience, is the key to scientific progress in this area."

Others think some fresh thinking is long overdue. "We have had a massive failure of the dominant economic model," says Eric Weinstein, a physicist working in mathematical finance for the Natron Group, a hedge fund in New York, "and we're trying to find the right people to deal with this failure. At least some of those people are likely to be unfamiliar voices and come from other parts of science."

At least some economists agree. The meltdown has shown that regulatory policies have to cope with far-from-equilibrium situations, says economist Blake LeBaron of Brandeis University in Waltham, Massachusetts. "Even fairly simple agent-based models can be used as thought experiments to see if there is something that hasn't been considered by the policy-makers."

LeBaron has spent the past decade and a half working with colleagues, including a number of physicists, to develop an agent-based model of the stock market. In this model, several hundred agents attempt to profit by buying and selling stock, basing their decisions on patterns they perceive in past stock movements. Because the agents can learn from and respond to emerging market behaviour, they often shift their strategies, leading other agents to change their behaviour in turn. As a result, prices don't settle down into a stable equilibrium, as standard economic theory predicts. Much as in the real stock market, the prices keep bouncing up and down erratically, driven by an ever-shifting ecology of strategies and behaviours.

Nor is the resemblance just qualitative, says LeBaron. Detailed analyses of the agent-based model show that it reproduces the statistical features of real markets, especially their susceptibility to sudden, large price movements. "Traditional models do not go very far in explaining these features," LeBaron says.

Another often-cited agent-based model got its start in the late 1990s, as the NASDAQ stock exchange in New York was planning to stop listing its stock prices as fractions such as 12¼ and instead list them as decimals. The goal was to improve the accuracy of stock prices, but the change would also allow prices to move by smaller increments, which could affect the strategies followed by brokers with unknown consequences for the market as a whole. So before making this risky change, NASDAQ chief Mike Brown hired BiosGroup, a company based in Santa Fe, New Mexico, to develop an agent-based model of the market to test the idea.

> **"We still implement new economic measures without any prior testing."**
> — Dirk Helbing

"Over ten years on the NASDAQ Board," says Brown, "I grew increasingly disappointed in our approach to studying the consequences of proposed market regulations, and wanted to try something different."

Once the model could reproduce price fluctuations in a mathematically accurate way, NASDAQ used it as a market wind tunnel. The tests revealed that if the stock exchange reduced its price increment too much, traders would be able to exploit strategies that would make them quick profits at the expense of overall market efficiency. Thus, when the exchange went ahead with the changeover in 2001, it was able to take steps to counter this vulnerability.

Agent-based models are also being used elsewhere in the private sector. For example, the consumer-products giant Proctor & Gamble of Cincinnati, Ohio, has used agent-based models to optimize the flow of goods through its network of suppliers, warehouses and stores. And Southwest Airlines of Dallas, Texas, has used agent-based models for routing cargo.

Despite such successes, however, financial regulators such as the US Securities and Exchange Commission (SEC) still don't use agent-based models as practical tools. "When the SEC changes trading rules, it typically has either flimsy or modest support from econometric evidence for the action, or else no empirical evidence and the change is driven by ideology," claims computational social scientist Rob Axtell of George Mason University in Fairfax, Virginia. "You have to wonder why Mike Brown is doing this, while the SEC isn't."

Risk of the new

A big part of the answer is that agent-based models remain at the fringe of mainstream economics, and most economists continue to prefer conventional mathematical models. Many of them argue that agent-based models haven't had the same level of testing.

Another problem is that an agent-based model of a market with many diverse players and a rich structure may contain many variable parameters. So even if its output matches reality, it's not always clear if this is because of careful tuning of those parameters, or because the model succeeds in capturing realistic system dynamics. That leads many economists and social scientists to wonder whether any such model can be trusted. But agent-based enthusiasts counter that conventional economic models also contain many tunable parameters and are therefore subject to the same criticism.

Familiarity wins out, notes Chester Spatt, former chief economist at the SEC. Regulators feel duty-bound to adhere to generally accepted and well-vetted techniques, he says. "It would be problematic for the rule-making process to use methods whose foundation or applicability were not established."

Still, agent-based techniques are beginning to enter the regulatory process. For example, decision-makers in Illinois and several other US states use computational models of complex electricity markets. They want to avoid a repeat of the disaster in California in 2000, when Enron and other companies, following market deregulation, were able to manipulate energy supplies and prices for enormous profit. Rich computational models have made it possible to test later market designs before putting them in place.

"We've had a lot of success in developing these models," says economist Leigh Tesfatsion of Iowa State University in Ames, who has led the development of an open-source agent-based model known as the AMES Wholesale Power Market Test Bed. "It has worked

© 2009 Macmillan Publishers Limited. All rights reserved

because we've focused on all the details of the real situation and can address questions that policy-makers really care about," she says.

Other models have successfully simulated financial markets. At Yale University, for example, economist John Geanakoplos, working with physicists Doyne Farmer of the Santa Fe Institute and Stefan Thurner of the Medical University of Vienna, has constructed an agent-based model exploring the systemic consequences of massive borrowing by hedge funds to finance their investments. In their simulations, the funds frequently get locked into a self-amplifying spiral of losses (see page 685) — much as real-world hedge funds did after August 2007.

> "Experts' complementary knowledge could 'collide', creating new knowledge."
> — Dirk Helbing

At the University of Genoa in Italy, meanwhile, Silvano Cincotti and his colleagues are creating an agent-based model of the entire European Union economy. Their model includes markets for consumer goods and financial assets, firms that interact with banks to obtain loans, and banks that compete with one another by offering different interest rates. Based on real economic data, the model currently represents some 10 million households, 100,000 firms and about 100 banks, all of which can learn and change their strategies if they find more profitable ways of doing business.

"We hope that these simulations will have an outstanding impact on the economic-policy capabilities of the European Union," says Cincotti, "and help design the best policies on an empirical basis."

This is the kind of ambition that has inspired Helbing. He doesn't pretend to be an economic modeller himself: since the early 1990s his own work has focused on simulations of human behaviour in relatively small groups — how traffic ebbs and flows through a road network, for example, or how crowds crush towards a door in a panic situation — as well as on experiments to test his predictions with real data. But that work has given Helbing a keen appreciation for the way complex collective phenomena can emerge from even the simplest individual interactions. If pedestrians can organize themselves into smoothly flowing streams just by trying to walk through a crowded shopping centre — as he has shown they do — just imagine how much richer the emergent phenomena must be in a group the size of a national economy.

Crisis logic

That observation acquired fresh force for Helbing after last year's global financial meltdown made it clear that a regulatory system based on conventional economic theory had failed.

"It's remarkable," he says, "that while any new technical device or medical drug has extensive testing for efficiency, reliability and safety before it ever hits the market, we still implement new economic measures without any prior testing."

To get around this impasse, he says, researchers need to reimagine the social and economic sciences on a larger scale. "I imagine experts from different fields meeting in one place for extended periods of time," he says, "so that their complementary knowledge could 'collide', creating new ideas, much as particle colliders create new kinds of particles." Ultimately, such an effort would bring together social scientists, economists, physicists, ecologists, computer scientists and engineers in a network of large centres for socioeconomic data mining and crisis forecasting, as well as in supercomputer centres for social simulation and wind-tunnel-like testing of policy.

That is a large ambition, Helbing admits — especially as he has only recently got tentative approval for a one-year grant from the European Commission to develop the idea. But now, in the aftermath of the meltdown, may be the time to start.

Axtell endorses that view. "Left to their own devices," he says, "academic macroeconomists will take a generation to make this transition. But if policy-makers demand better models, it can be accomplished much more quickly."

"The revolution has to begin here," agrees Weinstein, who helped organize a meeting in May at the Perimeter Institute for Theoretical Physics in Waterloo, Canada, that assembled the kind of interdisciplinary mix of experts that Helbing envisions. "And I think ideas from physics and other parts of science really have a chance to catalyse something remarkable." ■

Mark Buchanan is a science writer based in Cambridge, UK. After writing this story, he was involved in reviewing grant proposals on the topic of agent-based modelling.

See Editorial, page 667, and Opinion, pages 685 and 687.

© 2009 Macmillan Publishers Limited. All rights reserved

CORRESPONDENCE

Helping young scientists to speak for themselves

SIR — As you indicate in your Editorial 'Cheerleader or watchdog?' (*Nature* **459,** 1033; 2009), the quality of science journalism could be improved by better communication between scientists and the media. We should encourage this valuable skill in scientists at the outset.

I help an international team of high-school students to manage an online journal, *Young Scientists*, which is entirely written by people aged 12–20. To our knowledge, *Young Scientists* (www.ysjournal.com) is the only peer-reviewed science journal for school-aged students. Articles range from reviews of current hot topics to scholarly pieces of original research.

Many youngsters are now involved in scientific research, and at an increasingly early age — as demonstrated by the proliferation of science fairs around the world. Sadly, communication of all this promising work suffers because, once these bright young scientists have exhibited and gone home, their work goes with them. They need more opportunities to publish and share their ideas — before science journalists who are not scientists try to do their communication for them.

Science journalism is making increasing use of online media, which includes social networking sites. Who better to embrace it than our young scientists? If we can engender in them a critical perspective on the way science is reported and encourage them to participate in the process themselves, then we can look forward to a generation of scientists proficient at weighing up evidence and articulate in communicating it.

Christina Astin Physics Department, The King's School, Canterbury, Kent CT1 2ES, UK
e-mail: cma@kings-school.co.uk

Flu: no sign so far that the human pandemic is spread by pigs

SIR — Further to your Editorial 'Animal farm: pig in the middle' (*Nature* **459,** 889; 2009), the World Organisation for Animal Health (OIE) would like to clarify what is understood so far about how animals are associated with the human influenza A/H1N1 pandemic.

Although the human H1N1 virus contains gene sequences that have been identified in influenza viruses from swine, these are not present in exactly the same combination. The OIE has encouraged its members to intensify surveillance of pigs for infection, but there has been no evidence so far that swine are playing any role in the epidemiology or in the worldwide spread of the virus in the human population. It is likely that we shall never know the specific origin of this pandemic virus.

As you mention, the OIE has campaigned against calling the human disease 'swine flu'. Although the World Health Organization (WHO), the UN Food and Agriculture Organization (FAO) and the OIE have since agreed officially to rename the virus 'pandemic (H1N1) 2009', common use of the misleading term 'swine flu' is in danger of continuing. This initially prompted several countries to ban import of pigs and pig products or to destroy all their pig populations, without any benefit to public or animal health. It could cause further economic harm, in the same way that the H5N1 'avian flu' crisis of 2004 unnecessarily triggered a drop in people's consumption of poultry products. Such an unjustified disruption of trade would affect small farmers and animal producers around the world, more than a billion of whom are already living in poverty.

In 2005, the FAO and OIE set up a joint network of expertise on animal influenza. The network, OFFLU, was created to help the WHO obtain rapid access to circulating animal viruses for the early preparation of human vaccines. After the emergence of the pandemic virus in humans, OFFLU called for laboratories worldwide to aid public health by publicly sharing gene sequences of influenza virus identified in swine. As a result, it is proposed to expand the current OIE reference laboratories for avian influenza to cover all animal influenza viruses and to increase research on the behaviour of these viruses at the human–animal interface.

The OIE will continue to advise its members and the public on the control of potential zoonotic diseases, for example by strengthening veterinary infrastructure and stepping up surveillance and reporting capabilities in all countries, regardless of their trade potential.

Bernard Vallat World Organisation for Animal Health, 12 rue de Prony, 75017 Paris, France
e-mail: b.vallat@oie

Small but effective moves towards a greener China

SIR — Your Editorial 'Raising the standards' (*Nature* **459,** 1033–1034; 2009) reports on the pressure being imposed by non-governmental organizations on China's local governments to provide the public with more information about pollution. There is encouraging evidence that even a small organization can have an impact in this domain.

Ten years ago, there was hardly any environmental enforcement by civil society or by the markets in China. In 1999–2000, the World Bank collaborated on a pilot programme with the Chinese Academy of Environmental Planning, Nanjing University, the Zhenjiang Environmental Protection Bureau in Jiangsu Province and the Hohhot Academy of Environmental Sciences in Inner Mongolia. This experiment, aimed at publicizing information about environmental performance, was run in Hohhot and Zhenjiang. Although the programme was halted at the end of the pilot phase in Hohhot, it was sustained in Zhenjiang.

Despite the top leadership's intention to clean up China's environment, the evaluation system is biased towards economic development. A push from the bottom is badly needed to attract the attention of local governments to the environment.

The Pollution Information Transparency Index now has wide geographical coverage, and efforts are continuing by the Natural Resources Defense Council

© 2009 Macmillan Publishers Limited. All rights reserved

"Agents can be made to behave something like real people: prone to error, bias, fear and other foibles." Joshua M. Epstein

and the Institute of Public and Environmental Affairs in Beijing. So we have every reason to look forward to more informed public participation in environmental issues, stimulating local governments to embark on a path to a greener China.

Wanxin Li City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China, and Tsinghua Graduate School at Shenzhen, Shenzhen, China
e-mail: wanxili@cityu.edu.hk

Mystery ape: other fossils suggest that it's no mystery at all

SIR — Russell Ciochon, in his Essay 'The mystery ape of Pleistocene Asia' (*Nature* **459,** 910–911; 2009), makes passing reference to the Late Miocene ape *Lufengpithecus*, which is known from Lufeng in the Chinese province of Yunnan. Ciochon then immediately discounts the significance of *Lufengpithecus* because "the age was wrong". This assumption, however, leads up a blind alley.

Ciochon and his colleagues initially ascribed the teeth of a fossil found at Longgupo — in neighbouring Sichuan province — to *Homo* (W. Huang *et al. Nature* **378,** 275–278; 1995). Now he proposes a "mystery ape" to account for the Longgupo specimen and other similar material he recently observed in southern China.

He dismisses the possibility that these remains belong to descendants of *Lufengpithecus*. Yet it seems very likely that they do. The fauna recovered from Lufeng and Yuanmou, also in Yunnan — which have produced abundant fossils of *Lufengpithecus* — have also produced faunal remains directly ancestral to the *Stegodon–Ailuropoda* fauna of Pleistocene southern China (Z. Q. He and L. P. Jia (eds) *Yuanmou Hominoid Fauna;* Yunnan Science and Technology, 1997).

As both the Pleistocene apes *Gigantopithecus* and *Pongo* of southern China assuredly had Miocene antecedents, then so did Ciochon's mystery ape. Given their morphological and dimensional similarities, there is every reason to suspect that the mystery ape is none other than a descendant of *Lufengpithecus*, as originally proposed (for example, D. A. Etler *et al. Hum. Evol.* **16,** 1–12; 2001). Mystery solved.

Dennis A. Etler Anthropology Department, Cabrillo College, Aptos, California 95003, USA
e-mail: deetler@cabrillo.edu

Mystery ape: a call for taxonomic rigour

SIR — The Essay by Russell Ciochon on 'The mystery ape of Pleistocene Asia' (*Nature* **459,** 910–911; 2009) and the accompanying News story 'Early man becomes early ape' (*Nature* **459,** 899; 2009) announce that Ciochon has changed his mind about the taxonomic assignment of a 1.9-million-year-old hominoid partial jaw. But on what evidence is this reassignment based?

Whereas Ciochon and his colleagues originally considered the fossil on the *Homo* line (W. Huang *et al. Nature* **378,** 275–278; 1995), Ciochon now thinks it represents a "mystery ape" and that there is a group of them out there waiting to be discovered.

Although the News story included a photo and illustration of the fossil, I was unable to discern any evidence in either piece for taxonomic justification of the reassignment. I'm not a hominid expert so I'm not qualified to agree or disagree; I would just like to know if there are any anatomical characters — 'synapomorphies', in systematic parlance — that form the basis for this revised judgement, as one would expect for any taxon. If this is merely going with what other people thought, it is unclear why it is considered newsworthy.

Could one not certify what synapomorphies this fossil possesses, and place it at that particular node on the phylogenetic tree? Uncertain characters could then suggest further refinement if more information comes to light. How can one know that there was a "diversity" of Pleistocene mystery apes in southeast Asia without this kind of systematic rigour?

Kevin Padian Museum of Paleontology, University of California, Berkeley, California 94720, USA
e-mail: kpadian@berkeley.edu

Human uniqueness and the denial of death

SIR — Marc Hauser's Horizons article 'The possibility of impossible cultures' (*Nature* **460,** 190–196; 2009) carries an implicit assumption that cardinal aspects of human uniqueness arose by positive natural selection because they were beneficial to ancestral hominins. But this may not be the whole story.

Among key features of human uniqueness are full self-awareness and 'theory of mind', which enables inter-subjectivity — an understanding of the intentionality of others (see, for example, N. J. Emory and N. S. Clayton *Annu. Rev. Psychol.* **60,** 87–113; 2009). These attributes may have been positively selected because of their benefits to interpersonal communication, cooperative breeding, language and other critical human activities.

However, the late Danny Brower, a geneticist from the University of Arizona, suggested to me that the real question is why they should have emerged in only one species, despite millions of years of opportunity. Here, I attempt to communicate Brower's concept.

He explained that with full self-awareness and inter-subjectivity would also come awareness of death and mortality. Thus, far from being useful, the resulting overwhelming fear would be a dead-end evolutionary barrier, curbing activities and cognitive functions necessary for survival and reproductive fitness. Brower suggested that, although many species manifest features of self-awareness (including orangutans, chimpanzees, orcas, dolphins, elephants and perhaps magpies), the transition to a fully human-like phenotype was blocked for tens of millions of years of mammalian (and perhaps avian) evolution.

In his view, the only way these properties could become positively selected was if they emerged simultaneously with neural mechanisms for denying mortality. Although aspects such as denial of death and awareness of mortality have been discussed as contributing to human culture and behaviour (E. Becker *The Denial of Death;* Free Press, 1973), to my knowledge Brower's concept of a long-standing evolutionary barrier had not previously been entertained.

Brower's contrarian view could help modify and reinvigorate ongoing debates about the origins of human uniqueness and inter-subjectivity. It could also steer discussions of other uniquely human 'universals', such as the ability to hold false beliefs, existential angst, theories of after-life, religiosity, severity of grieving, importance of death rituals, risk-taking behaviour, panic attacks, suicide and martyrdom.

If this logic is correct, many warm-blooded species may have previously achieved complete self-awareness and inter-subjectivity, but then failed to survive because of the extremely negative immediate consequences. Perhaps we should be looking for the mechanisms (or loss of mechanisms) that allow us to delude ourselves and others about reality, even while realizing that both we and others are capable of such delusions and false beliefs.

Ajit Varki 9500 Gilman Drive, University of California, San Diego, La Jolla, California 92093-0687, USA
e-mail: a1varki@ucsd.edu

© 2009 Macmillan Publishers Limited. All rights reserved

OPINION

The economy needs agent-based modelling

The leaders of the world are flying the economy by the seat of their pants, say **J. Doyne Farmer** and **Duncan Foley**. There is, however, a better way to help guide financial policies.

In today's high-tech age, one naturally assumes that US President Barack Obama's economic team and its international counterparts are using sophisticated quantitative computer models to guide us out of the current economic crisis. They are not.

The best models they have are of two types, both with fatal flaws. Type one is econometric: empirical statistical models that are fitted to past data. These successfully forecast a few quarters ahead as long as things stay more or less the same, but fail in the face of great change. Type two goes by the name of 'dynamic stochastic general equilibrium'. These models assume a perfect world, and by their very nature rule out crises of the type we are experiencing now.

As a result, economic policy-makers are basing their decisions on common sense, and on anecdotal analogies to previous crises such as Japan's 'lost decade' or the Great Depression (see *Nature* **457,** 957; 2009). The leaders of the world are flying the economy by the seat of their pants.

P. NOBLE/REUTERS

Agent-based models could help to evaluate policies designed to foster economic recovery.

This is hard for most non-economists to believe. Aren't people on Wall Street using fancy mathematical models? Yes, but for a completely different purpose: modelling the potential profit and risk of individual trades. There is no attempt to assemble the pieces and understand the behaviour of the whole economic system.

There is a better way: agent-based models. An agent-based model is a computerized simulation of a number of decision-makers (agents) and institutions, which interact through prescribed rules. The agents can be as diverse as needed — from consumers to policy-makers and Wall Street professionals — and the institutional structure can include everything from banks to the government. Such models do not rely on the assumption that the economy will move towards a predetermined equilibrium state, as other models do. Instead, at any given time, each agent acts according to its current situation, the state of the world around it and the rules governing its behaviour. An individual consumer, for example, might decide whether to save or spend based on the rate of inflation, his or her current optimism about the future, and behavioural rules deduced from psychology experiments. The computer keeps track of the many agent interactions, to see what happens over time. Agent-based simulations can handle a far wider range of nonlinear behaviour than conventional equilibrium models. Policy-makers can thus simulate an artificial economy under different policy scenarios and quantitatively explore their consequences.

Why is this type of modelling not well-developed in economics? Because of historical choices made to address the complexity of the economy and the importance of human reasoning and adaptability.

The notion that financial economies are complex systems can be traced at least as far back as Adam Smith in the late 1700s. More recently John Maynard Keynes and his followers attempted to describe and quantify this complexity based on historical patterns. Keynesian economics enjoyed a heyday in the decades after the Second World War, but was forced out of the mainstream after failing a crucial test during the mid-seventies. The Keynesian predictions suggested that inflation could pull society out of a recession; that, as rising prices had historically stimulated supply, producers would respond to the rising prices seen under inflation by increasing production and hiring more workers. But when US policy-makers increased the money supply in an attempt to stimulate employment, it didn't work — they ended up with both high inflation and high unemployment, a miserable state called 'stagflation'. Robert Lucas and others argued in 1976 that Keynesian models had failed because they neglected the power of human learning and adaptation. Firms and workers learned that inflation is just inflation, and is not the same as a real rise in prices relative to wages.

Realistic behaviour

The cure for macroeconomic theory, however, may have been worse than the disease. During the last quarter of the twentieth century, 'rational expectations' emerged as the dominant paradigm in economics. This approach assumes that humans have perfect access to information and adapt instantly and rationally to new situations, maximizing their long-run personal advantage. Of course real people often act on the basis of overconfidence, fear and peer pressure — topics that behavioural economics is now addressing.

But there is a still larger problem. Even if rational expectations are a reasonable model of human behaviour, the mathematical machinery is cumbersome and requires drastic simplifications to get tractable results. The equilibrium models that were developed, such as those used by the US Federal Reserve, by necessity stripped away most of the structure of a real economy. There are no banks or derivatives, much less sub-prime mortgages or credit default swaps — these introduce too much nonlinearity and complexity for equilibrium methods to handle. When it comes to setting policy, the predictions of these models aren't even wrong, they are simply non-existent (see *Nature* **455,** 1181; 2008).

Agent-based models potentially present a way to model the financial economy as a complex system, as Keynes attempted to do, while taking human adaptation and learning into account, as Lucas advocated. Such models allow for the creation of a kind of virtual

© 2009 Macmillan Publishers Limited. All rights reserved


S. PLATT/GETTY IMAGES

universe, in which many players can act in complex — and realistic — ways. In some other areas of science, such as epidemiology or traffic control, agent-based models already help policy-making.

Promising efforts

There are some successful agent-based models of small portions of the economy. The models of the financial market built by Blake LeBaron of Brandeis University in Waltham, Massachusetts, for example, provide a plausible explanation for bubbles and crashes, reproducing liquidity crises and crashes that never appear in equilibrium models. Rob Axtell of George Mason University in Fairfax, Virginia, has devised firm dynamics models that simulate how companies grow and decline as workers move between them. These replicate the power-law distribution of company size that one sees in real life: a very few large firms, and a vast number of very small ones with only one or two employees.

Other promising efforts include the credit-sector model of Mauro Gallegati's group at the Marche Polytechnic University in Ancona, Italy, and the monetary model developed by Robert Clower of the University of South Carolina in Columbia and Peter Howitt of Brown University in Providence, Rhode Island. These models are very useful, but their creators would be the first to say that they provide only a tentative first step.

To see in more detail how an agent-based model works, consider the model that one of us (Farmer) has developed with Stefan Thurner of the University of Vienna and John Geanakoplos of Yale University to explore how leverage affects fluctuations in stock prices (published in a Santa Fe Institute working paper). Leverage, the investment of borrowed funds, is measured as the ratio of total assets owned to the wealth of the borrower; if a house is bought with a 20% down-payment the leverage is five. There are four types of agents in this model. 'Noise traders', who trade more or less at random, but are slightly biased toward driving prices towards a fundamental value; hedge funds, which hold a stock when it is under-priced and otherwise hold cash; investors who decide whether to invest in a hedge fund; and a bank that can lend money to the hedge funds, allowing them to buy more stock. Normally, the presence of the hedge funds damps volatility, pushing the stock price towards its fundamental value. But, to contain their risk, the banks cap leverage at a predetermined maximum value. If the price of the stock drops while a fund is fully leveraged, the fund's wealth plummets and its leverage increases; thus the fund has to sell stock to pay off part of its loan and keep within its leverage limit, selling into a falling market.

This agent-based model shows how the behaviour of the hedge funds amplifies price fluctuations, and in extreme cases causes crashes. The price statistics from this model look very much like reality. It shows that the standard ways banks attempt to reduce their own risk can create more risk for the whole system.

Previous models of leverage based on equilibrium theory showed qualitatively how leverage can lead to crashes, but they gave no quantitative information about how this affects the statistical properties of prices. The agent approach simulates complex and nonlinear behaviour that is so far intractable in equilibrium models. It could be made more realistic by adding more detailed information about the behaviour of real banks and funds, and this could shed light on many important questions. For example, does spreading risk across many financial institutions stabilize the financial system, or does it increase financial fragility? Better data on lending between banks and hedge funds would make it possible to model this accurately. What if the banks themselves borrow money and use leverage too, a process that played a key role in the current crisis? The model could be used to see how these banks might behave in an alternative regulatory environment.

> "The policy predictions of the models that are in use aren't wrong, they are simply non-existent."

Agent-based models are not a panacea. The major challenge lies in specifying how the agents behave and, in particular, in choosing the rules they use to make decisions. In many cases this is still done by common sense and guesswork, which is only sometimes sufficient to mimic real behaviour. An attempt to model all the details of a realistic problem can rapidly lead to a complicated simulation where it is difficult to determine what causes what. To make agent-based modelling useful we must proceed systematically, avoiding arbitrary assumptions, carefully grounding and testing each piece of the model against reality and introducing additional complexity only when it is needed. Done right, the agent-based method can provide an unprecedented understanding of the emergent properties of interacting parts in complex circumstances where intuition fails.

A thorough attempt to understand the whole economy through agent-based modelling will require integrating models of financial interactions with those of industrial production, real estate, government spending, taxes, business investment, foreign trade and investment, and with consumer behaviour. The resulting simulation could be used to evaluate the effectiveness of different approaches to economic stimulus, such as tax reductions versus public spending.

Holistic approach

Such economic models should be able to provide an alternative tool to give insight into how government policies could affect the broad characteristics of economic performance, by quantitatively exploring how the economy is likely to react under different scenarios. In principle it might even be possible to create an agent-based economic model capable of making useful forecasts of the real economy, although this is ambitious.

Creating a carefully crafted agent-based model of the whole economy is, like climate modelling, a huge undertaking. It requires close feedback between simulation, testing, data collection and the development of theory. This demands serious computing power and multidisciplinary collaboration among economists, computer scientists, psychologists, biologists and physical scientists with experience in large-scale modelling. A few million dollars — much less than 0.001% of the US financial stimulus package against the recession — would allow a serious start on such an effort.

Given the enormity of the stakes, such an approach is well worth trying. ■

J. Doyne Farmer is at the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA, and at LUISS Guido Carli in Rome, Italy, and founded the quantitative trading firm Prediction Company. **Duncan Foley** is Leo Model Professor of Economics at the New School for Social Research, 6 East 16th Street, New York 10003, USA, and an external professor at the Santa Fe Institute.
e-mails: jdf@santafe.edu; foleyd@newschool.edu

See Opinion, page 687, and Editorial, page 667.
Further reading accompanies this article online.

© 2009 Macmillan Publishers Limited. All rights reserved

Modelling to contain pandemics

Agent-based computational models can capture irrational behaviour, complex social networks and global scale — all essential in confronting H1N1, says **Joshua M. Epstein**.

As the world braces for an autumn wave of swine flu (H1N1), the relatively new technique of agent-based computational modelling is playing a central part in mapping the disease's possible spread, and designing policies for its mitigation.

Classical epidemic modelling, which began in the 1920s, was built on differential equations. These models assume that the population is perfectly mixed, with people moving from the susceptible pool, to the infected one, to the recovered (or dead) one. Within these pools, everyone is identical, and no one adapts their behaviour. A triumph of parsimony, this approach revealed the threshold nature of epidemics and explained 'herd immunity', where the immunity of a subpopulation can stifle outbreaks, protecting the entire herd.

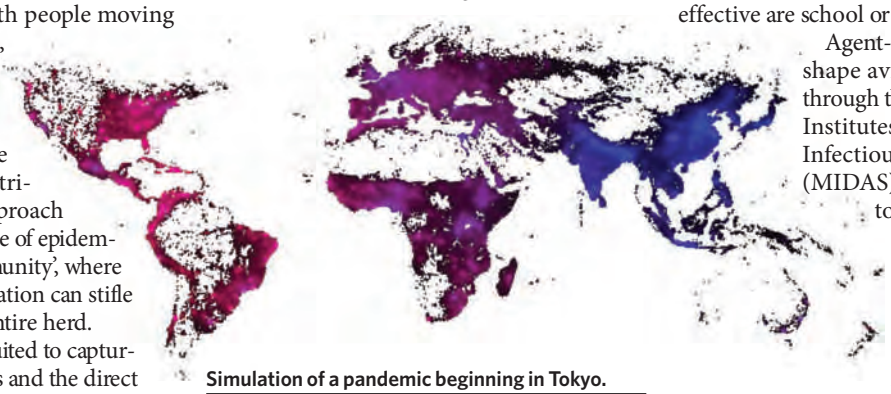
But such models are ill-suited to capturing complex social networks and the direct contacts between individuals, who adapt their behaviours — perhaps irrationally — based on disease prevalence.

Agent-based models (ABMs) embrace this complexity. ABMs are artificial societies: every single person (or 'agent') is represented as a distinct software individual. The computer model tracks each agent, 'her' contacts and her health status as she moves about virtual space — travelling to and from work, for instance. The models can be run thousands of times to build a robust statistical portrait comparable to epidemic data. ABMs can record exact chains of transmission from one individual to another. Perhaps most importantly, agents can be made to behave something like real people: prone to error, bias, fear and other foibles.

> "Agents can be made to behave something like real people: prone to error, bias, fear."

Such behaviours can have a huge effect on disease progression. What if significant numbers of Americans refuse H1N1 vaccine out of fear? Surveys and historical experience indicate that this is entirely possible, as is substantial absenteeism among health-care workers. Fear itself can be contagious. In 1994, hundreds of thousands of people fled the Indian city of Surat to escape pneumonic plague, although by World Health Organization criteria no cases were confirmed. The principal challenge for agent modelling is to represent such behavioural factors appropriately; the capacity to do so is improving through survey research, cognitive science, and quantitative historical study.

Robert Axtell and I published a full agent-based epidemic model[1] in 1996. Agents with diverse digital immune systems roamed a landscape, spreading disease. The model tracked dynamic epidemic networks, simple mechanisms of immune learning, and behavioural changes resulting from disease progression, all of which fed back to affect epidemic dynamics. However, the model was small (a few thousand agents) and behaviourally primitive.



Simulation of a pandemic beginning in Tokyo.

J. PARKER

Now, the cutting edge in performance is the Global-Scale Agent Model (GSAM)[2], developed by Jon Parker at the Brookings Institution's Center on Social and Economic Dynamics in Washington DC, which I direct. This includes 6.5 billion distinct agents, with movement and day-to-day local interactions modelled as available data allow. The epidemic plays out on a planetary map, colour-coded for the disease state of people in different regions — black for susceptible, red for infected, and blue for dead or recovered. The map pictured shows the state of affairs 4.5 months into a simulated pandemic beginning in Tokyo, based on a plausible H1N1 variant.

For the United States, the GSAM contains 300 million cyber-people and every hospital and staffed bed in the country. The National Center for the Study of Preparedness and Catastrophic Event Response at Johns Hopkins University in Baltimore is using the model to optimize emergency surge capacity in a pandemic, supported by the Department of Homeland Security.

Models, however, are not crystal balls and the simulation shown here is not a prediction. It is a 'base case' which by design is highly unrealistic, ignoring pharmaceuticals, quarantines, school closures and behavioural adaptations. It is nonetheless essential because, base case in hand, we can rerun the model to investigate the questions that health agencies face. What is the best way to allocate limited supplies of vaccine or antiviral drugs? How effective are school or work closures?

Agent-based models helped to shape avian flu (H5N1) policy, through the efforts of the National Institutes of Health's Models of Infectious Disease Agent Study (MIDAS) — a research network to which the Brookings Institution belongs. The GSAM was recently presented to officials from the Centers for Disease Control and Prevention in Atlanta, Georgia, and other agencies, and will be integral to MIDAS consulting on H1N1 and other emerging infectious diseases. In the wake of the 11 September terrorist attacks and anthrax attacks in 2001, ABMs played a similar part in designing containment strategies for smallpox.

These policy exercises highlight another important feature of agent models. Because they are rule-based, user-friendly and highly visual, they are natural tools for participatory modelling by teams — dinicians, public-health experts and modellers. The GSAM executes an entire US run in around ten minutes, fast enough for epidemic 'war games', giving decision-makers quick feedback on how interventions may play out. This speed may even permit the real-time streaming of surveillance data for disease tracking, akin to hurricane tracking. As H1N1 progresses, and new health challenges emerge, such agent-based modelling efforts will become increasingly important. ■

Joshua M. Epstein is director of the Center on Social and Economic Dynamics at the Brookings Institution, 1775 Massachusetts Avenue, Washington DC 20036, USA.
e-mail: jepstein@brookings.edu

1. Epstein, J. M. & Axtell, R. L. *Growing Artificial Societies: Social Science from the Bottom Up* Ch. V. (MIT Press, 1996).
2. Parker, J. A. *ACM Trans Model. Comput. S.* (in the press).

See Opinion, page 685, and Editorial, page 667.
Further reading accompanies this article online.

© 2009 Macmillan Publishers Limited. All rights reserved

BOOKS & ARTS

In Retrospect: Lamarck's treatise at 200

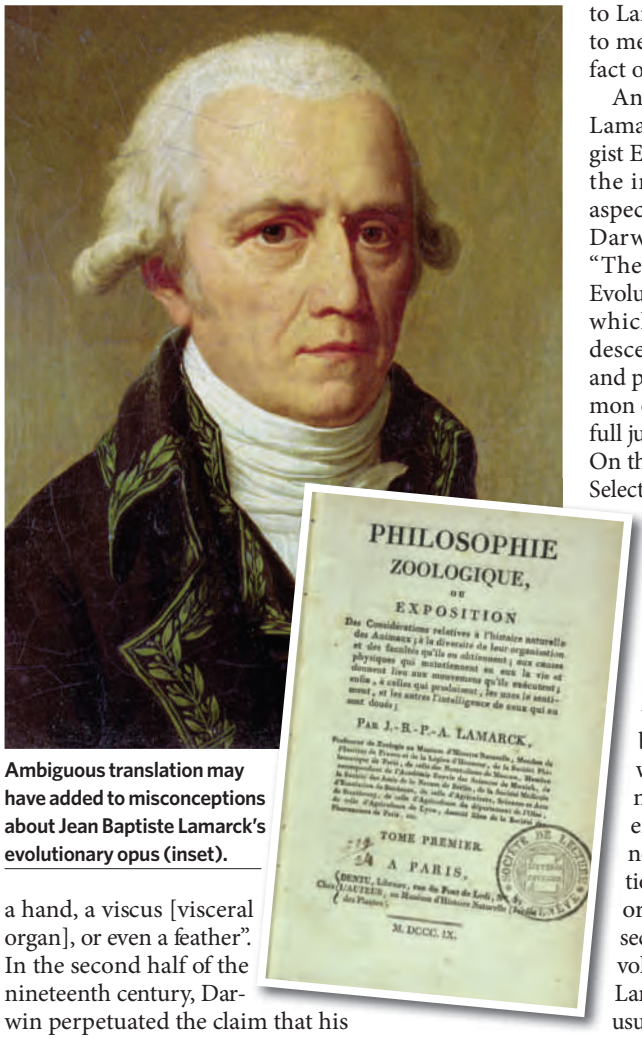
Fifty years before *On the Origin of Species*, a confusing, tiresome and prescient book laid the foundations of modern evolutionary theory, write **Dan Graur**, **Manolo Gouy** and **David Wool**.

Philosophie Zoologique (Zoological Philosophy)
by Jean Baptiste Lamarck
First published by the author: 1809.
Vol. I 428 pp; Vol. II 475 pp.
Translated by Hugh Elliot (Macmillan: 1914).
Vol. I translated by Ian Johnston: 1999
(http://tinyurl.com/philoszoo)

On 14 August 1809, Jean Baptiste Lamarck presented the two volumes of his most important book, *Philosophie Zoologique,* to France's Institut National des Sciences et Arts. Twenty years later, he died penniless, blind and scorned, surrounded by hundreds of unsold copies of his book. He was buried in a rented plot, exhumed and 'dispersed' five years later. Today, someone else occupies the grave of the man who founded the field of invertebrate zoology, coined the word biology and proposed the first scientific theory of evolution. His *Philosophie Zoologique* hasn't fared much better. It was mocked, ignored, belittled and purposely misunderstood for many years, remaining untranslated into English for 105 years. It took just 77 years, by contrast, to translate Charles Darwin's *On the Origin of Species* into Ukrainian.

But within the maddening, confusing and repetitive pages of Lamarck's exposition lurk concepts that are central to modern evolutionary thought. Stated in contemporary terminology, they include the ideas that species change through evolutionary time; that evolutionary change is slow and imperceptible; that evolution occurs through adaptation to the environment; that it generally progresses from the simple to the complex, although in a few cases it proceeds in reverse; and that species are related to one another by common descent. Furthermore, Lamarck incorporated into his theory the fact that the world is old, and proposed that the evolutionary process started with abiogenesis — the origin of life from inanimate matter.

So how and why has Lamarckism become a shorthand for foolishness? Lamarck's scientific reputation became tarnished soon after his death. In the 1830s, Georges Cuvier, Lamarck's fiercest opponent, published a 'eulogy' in French and English describing Lamarck's system as something that "cannot for a moment bear the scrutiny of anyone who has dissected



Ambiguous translation may have added to misconceptions about Jean Baptiste Lamarck's evolutionary opus (inset).

WELLCOME LIBRARY, LONDON; C. THEVENIN (1802–03)/PRIVATE COLLECTION/BRIDGEMAN ART LIBRARY

a hand, a viscus [visceral organ], or even a feather". In the second half of the nineteenth century, Darwin perpetuated the claim that his theory owed nothing to Lamarck's "nonsense". Later, Lamarck's name was damaged further by its association with Trofim Lysenko's quack genetics in the Stalinist Soviet Union. Recently, Lamarck has been invoked once more, again wrongly in our view, in the field of epigenetics — the study of phenotypic and gene-expression changes that occur without a change in the genetic material.

Lamarck did have a few fans. One was the great geologist and Darwin's friend Charles Lyell, who in his youth "devoured Lamarck" and late in life admitted having been unjust towards the French naturalist. Lyell felt that Darwin merely modified Lamarck's theory of evolution to coin his own, an attribution that greatly upset Darwin: "You often allude to Lamarck's work … it appeared to me extremely poor. I got not a fact or idea from it."

Another notable champion of Lamarck was the German biologist Ernst Haeckel. He recognized the injustice in attributing all aspects of evolutionary theory to Darwin, and in 1902 suggested: "The portion of the Theory of Evolution (*Entwickelungstheorie*), which maintains the common descent of all species of animals and plants from the simplest common original forms might … with full justice, be called Lamarckism. On the other hand, the Theory of Selection, or Breeding, might justly be called Darwinism."

Recognition of Lamarck's contribution is hindered by two persistent misconceptions. First, people wrongly assume that he believed in the direct induction of advantageous hereditary changes by the environment. Yet he writes repeatedly against this notion: "For, whatever the environment may do, it does not work any direct modification whatever in the shape and organization of animals." The second misconception concerns volition. A popular caricature of Lamarckism depicts an animal, usually a giraffe, wishing to reach the upper branches of trees, and acquiring a long neck through will alone. This error may have originated from the mistranslation of the French '*besoin*' — meaning 'need' — into the ambiguous term 'want', which can mean both 'desire' and 'need'. This poor choice by the 1914 translator was probably influenced by Cuvier's use of the word "*désir*" in his damning eulogy.

Of course, Lamarck did err. He believed in the inheritance of acquired characters (as did Darwin); adhered to the principle of plentitude — according to which any conceivable organism that can exist does exist; violently opposed Antoine Lavoisier and modern chemistry; and believed that science has a deistic purpose — similar to the accommodationism of modern biologists such as Ken Miller and Francis

© 2009 Macmillan Publishers Limited. All rights reserved

Collins. In fact, the amount of scientific rubbish that Lamarck put on paper certainly exceeds the quantity of good science in his scientific oeuvre. In this respect, he is no different from Aristotle, Isaac Newton, Darwin, Albert Einstein, Fred Hoyle or Francis Crick. But by writing about evolution directly rather than *en passant* (as did dozens of philosophers from Empedocles to Count Buffon), and by tackling the subject of evolution in scientific rather than poetical terms (as did Erasmus Darwin), Lamarck is without doubt the father of evolutionary theory.

In this year bracketed by two celebrations of Darwin — the 200th anniversary of his birth on 12 February and the sesquicentennial of the publication of his masterpiece on 24 November — let us pause on 14 August to ponder the man whose biological insight preceded *On the Origin of Species* by 50 years. ■

Dan Graur is John and Rebecca Moores Professor in the Department of Biology and Biochemistry at the University of Houston, Texas 77204-5001, USA; **Manolo Gouy** is directeur de recherche CNRS at the Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, 69622 Villeurbanne, France; and **David Wool** is professor emeritus of zoology at Tel-Aviv University, Tel Aviv 69978, Israel.
e-mail: dgraur@uh.edu

A passion for birds

Life List: A Woman's Quest for the World's Most Amazing Birds
by Olivia Gentile
Bloomsbury USA: 2009. 352 pp. $26, £25

If you had less than one year left to live, how would you spend your days? After being diagnosed with terminal cancer, Phoebe Snetsinger, the subject of Olivia Gentile's first book, invested her time trying to see every bird species in the world. In the process, this American grandmother became the first person ever to see 8,000 species of birds. *Life List* is her story.

Birdwatching is typically dismissed as a quiet hobby pursued by eccentrics, but it can be more like an extreme sport. Most birders keep a record of all the species they've spotted — their 'Life List' — and its size is a source of prestige. Intense competition results. It is a pastime often dominated by middle-aged men who seek out globe-trotting, cliff-dangling adventures, punctuated by bouts of dysentery and malaria, to fulfil their quest to see the rarest birds in the world.

Birders share attributes with many scientists who may not know where the line between passion and obsession lies. But obsession requires extreme sacrifices.

Phoebe didn't start out noticing birds. In her youth, she was a tomboy who distinguished herself as a gifted student with a natural affinity for writing, languages and the sciences. But in the 1950s, young women's futures were limited, so Phoebe followed the expected path: marriage and children. But dedication to her family did not relieve the boredom, frustration and intellectual starvation that accompanied suburban life. Depression set in.

One sunny day, a neighbour took Phoebe into the back yard, put a pair of binoculars into her hands and pointed to a small bird perched in a treetop. From the moment she set eyes upon the blazing orange throat of that Blackburnian Warbler (*Dendroica fusca*), she was hooked. She purchased binoculars, studied field guides and went out birding with her neighbour several times a week. Her remarkable memory and enthusiasm overcame her innate shyness, so she quickly befriended other birders. Birdwatching became Phoebe's freedom from the cage of domesticity.

As her skills improved, Phoebe began travelling farther to see birds. But everything changed in 1981, just a few months short of her fiftieth birthday, when she was diagnosed with a malignant melanoma. She was given less than one year to live. At roughly the same time, she received an inheritance from the estate of her father, multimillionaire Leo Burnett, who had died ten years previously. With the blessings of her husband and children, Phoebe used her inheritance to pursue her passion. She set out to see more birds than anyone else had ever done before.

COURTESY OF THE SNETSINGER FAMILY

Phoebe Snetsinger logged more than 8,000 bird species.

Despite her diagnosis, Phoebe did not die from cancer. She spent the next 18 years pursuing birds into exotic places, through war-torn lands, despite several injuries and the death of a birding companion. She persisted even after being assaulted in New Guinea. But her decision to pursue birds meant sacrifices elsewhere. It often took her away from family events: Phoebe missed weddings, funerals and christenings. Eventually her marriage was at stake.

All this ended abruptly in Madagascar in 1999: Phoebe was killed when the van she was travelling in overturned. She had just seen a rare species of vanga, a stunning bird that had only recently been described.

Life List is riveting and, like its subject, demonstrates a passion bordering on obsession. The index is extensive and there are detailed chapter notes, citing interviews with Phoebe's family and friends, referencing scientific papers, magazine articles and books, including Phoebe's personal memoir, *Birding on Borrowed Time* (American Birding Association, 2003).

Yet the story of a suburban housewife and mother-of-four who became a legend in the testosterone-driven world of competitive birding is more than a biography. It raises themes that echo through all our lives, from the restriction of people's roles by society, to questions of how best to spend one's days on Earth. Is pursuing a rare bird a trivial pursuit, or a chase worthy of respect? Ultimately, *Life List* asks what it means to live, and die, well. ■

Devorah Bennu is a researcher, ornithologist and writer who writes the blog 'Living the Scientific Life (Scientist, Interrupted)' under the pseudonym GrrlScientist.
e-mail: grrlscientist@gmail.com

© 2009 Macmillan Publishers Limited. All rights reserved

NATURE|Vol 460|6 August 2009

Playing the con game of academe

Lives in Science: How Institutions Affect Academic Careers
by Joseph C. Hermanowicz
University of Chicago Press: 2009.
344 pp. $45

Young scientists often aim for research-focused professorships at elite universities. Those who achieve this goal spend their lives tirelessly working towards the next great finding, hopeful of recognition from their scientific peers. Other graduates take academic jobs at less-prestigious colleges and universities, where they may divide their careers between research, service, teaching and outside pursuits. Who is happier in the end?

In *Lives in Science*, sociologist Joseph Hermanowicz examines the career paths and overall satisfaction of a small group of physicists using an innovative longitudinal research method. In 1994, he interviewed 60 randomly selected tenured and tenure-track physicists at six universities across the United States. The results were published in his 1998 book *The Stars Are Not Enough: Scientists — Their Passions and Professions*. Ten years later, Hermanowicz interviewed the same scientists again to assess how their careers had changed in the interim. He recorded their number of publications, career focus, aspirations, orientation to work and focus on work as opposed to family.

Hermanowicz likens academic careers to a 'con game' in which faculty members are victims. The process starts in graduate school, where students in all academic disciplines are reverentially taught about the exemplary research of those who have excelled in the field. They learn to emulate the pantheon of great researchers. Thus, all start out expecting to achieve greatness; but few do so. Robbed of the status and recognition they once sought, physicists at less-prestigious universities must learn to console themselves.

The scientists are compared by age. In his 1998 analysis, Hermanowicz defined three groups: 'early career' included those who earned PhDs after 1980; 'middle career' scientists earned theirs between 1970 and 1980; and 'late-career' scientists were PhDs before 1970. Ten years later, the physicists had all moved up a group, with the late-career group moving into a new category — 'post career', consisting of those near or at retirement.

The physicists are also categorized into three groups based on their workplaces: 'elite' research universities, 'pluralist' universities that "place a premium on both research and teaching" and 'communitarian' universities that emphasize "teaching in the presence of research". Hermanowicz, like many qualitative researchers, allows the respondents to define their own institution types, which may differ from the more objective rankings of the US National Research Council. This allows him to identify overlapping categories, such as individual scientists who behave like members of an elite while working at a pluralist university.

T. STEWART/CORBIS

Scientists who teach at less-prestigious universities are more satisfied at retirement than those at elite institutions because they had more realistic career expectations.

Hermanowicz finds that a faculty member's level of career satisfaction at retirement depends on the prestige of their institution and the scientific reputation they are able to achieve. Those at less-prestigious universities, who were also more likely to have graduated from similar institutions, were generally satisfied because of the balance they ultimately achieved in their lives. Like other academics, they had once hoped to achieve scientific greatness, but quickly realized that such recognition would elude them. They dealt with disappointment about their career paths early on.

> **"Students in academic disciplines all start out expecting to achieve greatness; but few do."**

By contrast, physicists who got the early prize of an elite university job were satisfied with their careers — until the end. Then they were hit with the realization that the scientific recognition for which they had striven so long would now go to younger scientists. For the first time, this elite group's "expectations for their careers exceed reality" and their satisfaction was low.

The study is grounded in the theories and methods of sociology, which may be unfamiliar to readers from the natural sciences. Hermanowicz bases his comparison on Erving Goffman's 1952 analysis of how individuals reconcile their expectations, which are socially produced, with the reality that society limits the ability to achieve those expectations. Also applied is the concept of anomie, first outlined by the French sociologist Emile Durkheim in the late nineteenth century: a loss of societal norms that results from people's expectations for the future exceeding the realities of life. People can experience the disillusionment of anomie in various aspects of their lives, but for scientists it occurs when the desire to achieve recognition from one's peers exceeds the opportunities to do so.

Hermanowicz makes many assumptions. He assumes, as do the physicists in his study, that research is the highest form of scholarly endeavour. He refers to teaching as an undesirable activity — as "acceptable unproductivity". In focusing solely on academics, Hermanowicz ignores the vast numbers of physicists who work outside academia. If it is the ability to achieve scientific recognition that delivers career satisfaction, as he argues, then perceptions might evolve differently in government labs and in private industry. These researchers, too, would once have expected to join the scientific pantheon but have taken other roles. How have they adjusted their expectations to more restricted research avenues?

Lives in Science reveals that all scientists are socially conditioned to contribute substantially to the knowledge base and expect to receive recognition for it. But all must reconcile themselves to the shortcomings of the academic game. With research pressure growing in less-prestigious universities, and with limited resources, anomie will remain with us. Its cure is to require graduate institutions to present a more realistic picture of what it means to be a scientist. ■

Rachel Ivie is a sociologist and assistant director of the Statistical Research Center at the American Institute of Physics, 1 Physics Ellipse, College Park, Maryland 20740, USA.
e-mail: rivie@aip.org

© 2009 Macmillan Publishers Limited. All rights reserved

Me, environmentalist

Tarzan! Or Rousseau Among the Waziri
Quai Branly Museum, Paris, France
Until 27 September

Summer visitors to Paris might hear Tarzan's distinctive roar above the sound of traffic. The subject of an exhibition at the city's Quai Branly Museum (Musée du Quai Branly), the fictional loin-clothed hero provides possibly the most oblique take on Darwinism seen during the bicentennial year.

Tarzan! Or Rousseau Among the Waziri uses this cultural icon to examine depictions of the African jungle and the relationships between humans, apes and other animals in it. What is surprising, on rereading *Tarzan of the Apes* almost a century after its first appearance in the October 1912 issue of US magazine *The All-Story*, is author Edgar Rice Burroughs's evident interest in nature versus nurture. He studied *On the Origin of Species* closely.

"I was mainly playing with the idea of a contest between heredity and environment," Burroughs wrote in *Writer's Digest* in 1932. He selected an infant child "at an age at which he could not have been influenced by association with creatures of his own kind" and placed him in "an environment as diametrically opposite that to which he had been born as I might well conceive".

Early ecowarrior? The exhibition examines film depictions of the African jungle and its inhabitants through Tarzan.

Tarzan was instantly popular. The tale of this aristocratic hero, the orphaned son of John Clayton (Lord Greystoke) and his wife Lady Alice, raised by the female ape Kala whose own infant had been killed, retains its appeal. Burroughs turned his writing into a business, producing 22 Tarzan adventures that led to 42 feature films, 15,000 comic books, and innumerable cartoons and television series.

It is surprising that France's national museum of non-European indigenous art, cultures and civilizations, with its enormous collection of African, Asian, Oceanic and American artefacts, has devoted an exhibition to Tarzan. "Our exhibition allows the public to discover how Tarzan, an icon of popular culture, was created and to decipher the heroic myth that he embodies," says exhibition curator Roger Boulay, an ethnologist.

The subtitle 'Rousseau Among the Waziri' links Tarzan to the eighteenth-century philosopher Jean-Jacques Rousseau's ideas on human perfectibility, in evolving from a primitive natural state to civilized society. The Paris exhibition reminds us that Tarzan was instructed in etiquette and was taught to speak French by Lieutenant Paul d'Arnot, the naval officer he had rescued from African cannibals. Later, as they walked up the coast together for four weeks to reach a river port, Tarzan metamorphosed under d'Arnot's instruction into the white-clad Monsieur Tarzan, unfazed by cutlery and dinner conversation.

The exhibits include ethnographic objects and stuffed African animals from several French museums; as well as original comic strips and film and television clips, which fuelled the public's vision of a fantasy Africa. Juxtaposing the comic-book imagery with less familiar African artefacts shows that our understanding of Tarzan owes as much to drawings by Tarzan illustrators Burne Hogarth and Joe Kubert, and to film actors including Frank Merrill and Johnny Weissmuller, as it does to Burroughs' text.

The exhibition redefines Tarzan as a twentieth-century ecowarrior, ahead of his time in fighting animal poachers and slave traffickers. It encourages us to look beyond clichés to what it is to be human; part of the animal kingdom but with the capacity to reason — and to take considerate actions. ■

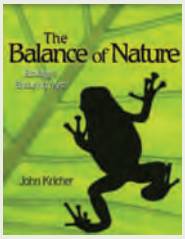
Colin Martin is a writer based in London.
e-mail: cmpubrel@aol.com

© 1936 MGM, ALL RIGHTS RESERVED/© TARZAN (TM) AND EDGAR RICE BURROUGHS (TM) OWNED BY EDGAR RICE BURROUGHS, INC. AND USED BY PERMISSION/ARCHIVES STANISLAS CHOKO

ECOLOGY READING

In *Notebooks from New Guinea* (Oxford University Press, 2009), tropical biologist Vojtech Novotny describes vividly what it is like to work deep in the malaria-infested Papuan rainforest. Sharing his personal experiences of setting up a research station in this remote and lawless place, he reflects on the clash between the cultures of Papua New Guinea and Europe. Novotny is humbled by the folk knowledge of local tribes, and colourfully describes their customs and interactions with loggers and conservationists.

The urban wilderness is the home of naturalist Lyanda Lynn Haupt. In *Crow Planet* (Little, Brown, 2009), she weaves into a series of stories the science, history and mythology of these independent and charismatic birds. Crows are noisy, boisterous and quick to learn and adapt to urban life. Through her closely observed portraits of these creatures, Haupt urges us to pay more attention to nature within our city landscapes.



John Kricher tackles the history of ecology in his new book, *The Balance of Nature* (Princeton University Press, 2009). Arguing that nature is far from poised and is constantly in flux, he asks why we hold on to the idea of ecological balance. He finds that the roots of our desire for order are ancient, and predate Aristotle. But solutions to today's environmental problems require us to take a more dynamic view of nature. Ultimately, he explains, evolution is the driver of natural systems. **J. B.**

© 2009 Macmillan Publishers Limited. All rights reserved

NEWS & VIEWS

DEMOGRAPHY

Babies make a comeback

Shripad Tuljapurkar

The population of some wealthy countries is shrinking because of a declining birth rate. It comes as a surprise, and one with policy implications, that after a certain point of development that trend can reverse.

In many industrialized nations, including Japan, South Korea, Germany and Italy, and much of southern and eastern Europe, fertility is far below replacement — the level at which enough children are born to replace their parents. Many of these countries do not accept (or want) immigrants to make up this deficit, so their populations are projected to decline over the next 25–50 years, with potentially scary consequences[1]. Low fertility means that women delay and reduce childbearing[2], choices that typically follow improvements in education, wealth and health. Widespread fertility decline and its associated problems have seemed inevitable. But on page 741 of this issue[3], Myrskylä, Kohler and Billari brighten this prospect by presenting evidence that such declines may be expected to reverse.

For many years, environmental concerns have been used to argue that it would be a good thing if human populations became smaller. If that's true, should we not welcome low fertility and population decline wherever it occurs? The difficulty with this view is that although smaller populations may indeed be desirable in the long term, in the short term population decline poses challenges that we do not know how to manage. Low fertility means fewer babies, and eventually a smaller workforce that would have to pay higher per capita costs of infrastructure and social support systems. A consequence of low fertility and long lives is an ageing population with its attendant social and economic effects. National economic output would probably decline along with the size of the workforce. Political and military capability and influence would decline along with population. Thus, for many rich countries, population decline is a serious concern.

Myrskylä *et al.*[3] examine the relationship between fertility and the human development index (HDI), a measure of education, income and lifespan[4]. Fertility decreases with increasing HDI during early stages of development. But at high levels of development, fertility in many countries increases with HDI. This is the first evidence that fertility levels might move back towards the replacement level in such countries as Italy, Spain, the Netherlands, Germany and Sweden. Perhaps babies will be 'in' again in the richest countries.

B. HARRINGTON/PHOTOLIBRARY

Bundles of joy — and future taxpayers.

To understand the fuss about low fertility and population decline, consider how fertility affects population number. Annual fertility is measured as birth rate expected according to a woman's age, and it is summarized by the total fertility rate (TFR), the total number of children a woman could have at those age-specific rates. Replacement fertility in countries with long life expectancies is at a TFR of about 2.1. Many rich countries now have fertilities far below this replacement level, close to the record low fertilities of Spain, Japan and Italy, which had TFRs close to 1.3 in 2005. If fertility stayed at that level, the populations of these countries would eventually decline at about 1.5% per year. Annual immigration at that amount would just offset the decline, but would also lead to a rapid increase in the number of foreign-born residents. The latter factor comes with political concerns about the economic, social and cultural assimilation of immigrants, concerns that remain even in the United States, a country with a long history of immigration at or near such levels.

To obtain the HDI, education, income and length of life are evaluated relative to best-possible values and combined into a score on a scale of 0 to 1. Most low-fertility countries, including Spain and Italy, had HDI levels of more than 0.9 in 2005; more broadly, the HDI in most countries has been increasing over time. Myrskylä *et al.* discovered that in virtually all countries, the TFR falls as the HDI increases up to about 0.86. By contrast, when the HDI rises above that level, the TFR increases in many (but not all) countries.

The authors argue that high HDI levels (above 0.86) may result in changes that benefit women and make it easier for them to choose to have children. Increases in development in rich countries come about as a result of higher educational attainment for women, increases in the percentage of women in the labour force, and an increase in women's incomes. These changes make it likely that women, and couples, will find it easier to pay the high economic price of having children. In addition, women with more skills and work experience — important elements of what we call human capital — will probably find it easier to move out of jobs to have children and then to move back into jobs once the children are in school.

How far can these results go in alleviating concerns over population decline? An increase of 0.01 in the HDI can increase the TFR by 0.03

© 2009 Macmillan Publishers Limited. All rights reserved

NATURE|Vol 460|6 August 2009

or more[3], or equivalently can increase the eventual annual growth rate by about 0.06%. This may seem like considerable leverage. But the HDI can't go above 1, and many low-fertility countries already have HDI levels of around 0.93. Therefore, the best one can expect is an increase in TFR of 0.2, which is equivalent to raising the growth rate by about 0.4% from its currently projected lows. Countries such as Spain or Italy would still be below replacement, although, setting the social and political considerations to one side, they would be able to maintain their populations with far fewer immigrants.

Myrskylä and colleagues[3] find important exceptions to the relationship between the TFR and HDI — in some countries, including Japan, South Korea and Canada, the TFR continued to fall even when the HDI rose above 0.86. What might be happening here? The authors suggest that the positive effects of increasing HDI on women's decisions to have children may not apply in Asian countries because of social or cultural characteristics. Perhaps so, but what about Canada? These puzzling findings may instead be due to use of the HDI, which does not directly tell us which aspects of human development affect women rather than men. A different measure, the gender development index (GDI)[4], describes the difference between male and female development. It would be useful to examine the relationship between the TFR and GDI, and to ask if countries such as Japan or Canada have a noticeable difference between trends in the HDI and GDI.

A final point worth stressing is that Myrskylä and colleagues[3] also show that fertility in developing countries, which have HDI levels much below 0.86, falls with increasing development. The social and environmental challenges of burgeoning populations in many developing countries, such as Bangladesh, Egypt, India and Pakistan, can only be addressed if these countries achieve and maintain low, below-replacement, fertility. Even in China, where low fertility was achieved by fiat, the maintenance of low fertility must eventually be driven by individual choice. In these and other developing countries, increased human development, especially development that benefits women, is still the most powerful and most democratic route to achieving and maintaining lower populations. ■

Shripad Tuljapurkar is in the Stanford Center for Population Research, and the Department of Biological Sciences, Stanford University, Stanford, California 94305-5020, USA.
e-mail: tulja@stanford.edu

1. United Nations Population Division. *Replacement Migration: Is it a Solution to Declining and Ageing Populations?* (United Nations, 2002).
2. Lesthaeghe, R. & Willems, P. *Pop. Dev. Rev.* **25,** 211–228 (1999).
3. Myrskylä, M., Kohler, H.-P. & Billari, F. C. *Nature* **460,** 741–743 (2009).
4. United Nations Development Programme. *Statistics of the Human Development Report* http://hdr.undp.org/en/statistics/indices (2008).

GALAXY FORMATION

Too small to ignore

Karl Glazebrook

A study of one galaxy's dynamics backs up previous claims that surprisingly compact galaxies existed in the early Universe. But how such objects blew up in size to form present-day galaxies remains a puzzle.

Giant red elliptical galaxies are the oldest and most massive assemblies of stars in the nearby Universe. Large optical telescopes have tracked their evolution back through 11 billion years — about 80% of the Universe's lifetime — by observing them at large cosmological redshifts[1–3]. Observations seemed to indicate that nothing much had happened to them over this time, except that they grew rarer in the more distant past. It was thus a surprise when astronomers discovered recently that these galaxies have grown in size by a factor of five over this period while barely changing in mass. This is like suddenly discovering that Roman Londinium had the same population as Greater London does today. This extreme size and density evolution was not predicted by theories of galaxy formation, and remains difficult to explain.

One possibility is that measurements of the galaxies' masses that are based on their luminosities are flawed, not only because they might be missing starlight but also because they are not sensitive to the galaxies' invisible dark-matter component (which does not contribute significantly to mass in the luminous regions, at least for nearby ellipticals). On page 717 of this issue, van Dokkum and colleagues[4] report the first 'dynamical' mass measurement — which is sensitive to both the visible and the dark-matter component — for an individual red compact galaxy, known as 1255–0, that is seen 10.7 billion years ago and is less than 1 kiloparsec (~3,000 light years) in size. This galaxy is about four times more massive and five to six times smaller than the Milky Way spiral. Distant red compact objects are widely considered to be ancestors of today's ellipticals owing to their similar stellar populations and morphologies. However, nearby ellipticals of similar mass have sizes of 3–10 kpc; none is of similar size and mass to 1255–0 (ref. 5).

A dynamical mass measurement is definitive but requires resolution of the internal velocity structure of the galaxy. If high-redshift ellipticals really are small but massive, it follows from Newton's law that their stars' velocities should be very high. Because stars in ellipticals generally move on eccentric orbits, this is measured using the stellar-velocity dispersion, which quantifies the average spread of velocities and can be determined from the subtle Doppler broadening of absorption lines in the galaxies' spectra.

Measuring the velocity dispersion for 1255–0 was a tour de force[4]. Most of the light of such a high-redshift galaxy is redshifted into the near-infrared waveband (1–2 μm), where the airglow emission from the night sky contributes an enormous, noisy background. Van Dokkum *et al.*[4] sustained a heroic 29-hour exposure time using the 8-metre Gemini South Telescope's Near-Infrared Spectrograph. This was just enough to detect the absorption lines in the galaxy's spectrum — a feat in itself for such a high-redshift galaxy — and to measure the velocity dispersion. The measurement was partly helped by the large value of the final result: 510 km s^{-1}, the largest ever measured for any galaxy, but a value it had to have if it was as massive and as compact as previous measurements of its stellar mass had suggested[6].

Van Dokkum and colleagues' result[4] is especially surprising because an earlier velocity-dispersion measurement[7], based on a composite spectrum of several distant ellipticals (an interesting technique, albeit subject to its own problems), had measured a much lower value, sparking debate about the reliability of the stellar-mass measurements and the role of dark matter. The new measurement is significant because it is for a single object and so implies unequivocally that 1255–0 does have a density much higher than any nearby galaxy.

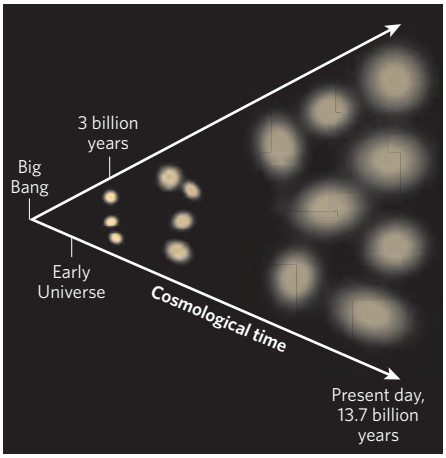


Figure 1 | Bloating galaxies. As the Universe evolves with time, elliptical galaxies, which are believed to be very compact at early epochs, maintain much the same mass but become bigger, more diffuse and more abundant. Newly produced ellipticals somehow have similar properties to the evolving, older ones. Van Dokkum and colleagues' spectroscopic analysis[4] of a galaxy seen when the Universe was about three billion years old gives clear-cut evidence of the compactness and high density of such a high-redshift object.

© 2009 Macmillan Publishers Limited. All rights reserved

So why are these compact, high-redshift galaxies such a theoretical conundrum? Elliptical galaxies have long been seen as the endpoint of galaxy formation: when a star-forming spiral or irregularly-shaped galaxy, full of young blue stars, has its star formation quenched by some astrophysical-feedback process, it quickly ages to become a 'red and dead' elliptical. As the overall star-formation rate of the Universe winds down with time, it seems natural to find an increasing number of these galaxy fossils full of old red stars. But such red galaxies are found to be more massive than any star-forming galaxy, so something extra is needed to provide the mass.

The consensus has been that elliptical galaxies have also assembled through mergers of smaller galaxies, a process naturally expected in current galaxy-formation theories[8]. The most massive ellipticals would be the result of major mergers of smaller ellipticals — with these progenitors having been of roughly equal mass. Elliptical galaxies are observed to follow tight scaling relationships between size, mass and velocity, which one might think would be seriously disturbed by mergers. However, computer simulations show that mergers simply 'move' the galaxies along the relationships without making them significantly less tight[9].

But this picture breaks down when size evolution is taken into account: if you merge enough elliptical galaxies at high redshift to account for the size change, you also make many more high-mass galaxies than are observed in the nearby Universe. An alternative is that if mergers are predominantly minor — those in which a low-mass object merges into one of much larger mass — size growth can be achieved without a substantial increase in mass[10]. However, low-mass galaxies generally contain a lot of young stars, so this seems inconsistent with the observed old stellar populations of the high-redshift compact galaxies and their nearby descendants.

This 'lack of fit' with the standard picture of elliptical-galaxy formation has driven a search for ways other than mergers by which the size of these galaxies could have blown up. For example, feedback processes such as an energy injection from a supernova[11] or quasar[12] could achieve that by expelling gas from the galaxy slowly (or rapidly), making the galaxy's gravitational potential well shallower and moving stars into larger orbits. But these processes require a level of star or quasar activity that has not been observed. A more exotic explanation could involve the yet unknown nature of dark matter.

Any successful explanation of the size evolution must solve what I call the synchronization problem, which in my view is the most fundamental. The size–mass scaling relationship is tight in the nearby Universe, and possibly also at high redshift. It is just the normalization of this relationship that evolves. There are no massive compact elliptical galaxies today. Therefore, the high-redshift (early-epoch) compact galaxies must be growing in size with time (Fig. 1). But, at the same time, the Universe is making new elliptical galaxies, and somehow both the growing and the newly formed galaxies fall within the same tight size–mass relationship at all epochs. Their evolution is 'synchronized' through some process that is either a coincidence or an important new piece of astrophysics.

A lot hinges on the interpretation of van Dokkum and colleagues' single velocity-dispersion measurement[4] of 1255–0. As large telescopes acquire new multi-object, near-infrared spectrographs, we can expect to see many hundreds of such velocity-dispersion measurements in the next few years. We can also expect to see improved measurements of the structural and environmental properties of these compact galaxies, which will help us to figure out how bad the problems we have in explaining these objects really are. It remains to be seen whether we need conventional or novel explanations for their astounding growth into the most massive elliptical galaxies we see today. ■

Karl Glazebrook is at the Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia.
e-mail: karl@astro.swin.edu.au

1. Abraham, R. G. *et al. Astrophys. J.* **669,** 184–201 (2007).
2. Cimatti, A. *et al. Nature* **430,** 184–187 (2004).
3. van Dokkum, P. G. *et al. Astrophys. J.* **638,** L59–L62 (2006).
4. van Dokkum, P. G., Kriek, M. & Franx, M. *Nature* **460,** 717–719 (2009).
5. Trujillo, I. *et al. Astrophys. J.* **692,** L118–L122 (2009).
6. Kriek, M. *et al. Astrophys. J.* **700,** 221–231 (2009).
7. Cenarro, A. J. & Trujillo, I. *Astrophys. J.* **696,** L43–L47 (2009).
8. De Lucia, G., Springel, V., White, S. D. M., Croton, D. & Kauffmann, G. *Mon. Not. R. Astron. Soc.* **366,** 499–509 (2006).
9. Boylan-Kolchin, M., Ma, C.-P. & Quataert, E. *Mon. Not. R. Astron. Soc.* **369,** 1081–1089 (2006).
10. Bezanson, R. *et al. Astrophys. J.* **697,** 1290–1298 (2009).
11. Damjanov, I. *et al. Astrophys. J.* **695,** 101–115 (2009).
12. Fan, L., Lapi, A., De Zotti, G. & Danese, L. *Astrophys. J.* **689,** L101–L104 (2008).

ARCHAEOLOGY

The earliest musical tradition

Daniel S. Adler

Music is a ubiquitous element in our daily lives, and was probably just as important to our early ancestors. Fragments of ancient flutes reveal that music was well established in Europe by about 40,000 years ago.

The Palaeolithic caves of the Swabian Jura in southwestern Germany have been a source of valuable and often provocative archaeological discoveries for many decades. In particular, finds of figurative art from the early Aurignacian — the earliest Upper Palaeolithic archaeological culture associated with modern humans in Europe — suggest that these hunter–gatherers had the knowledge, expertise, incentive and time to craft sophisticated objects for use in ritual activities. These activities probably served to affirm group affiliation, signal social identity and mark important social events or rites of passage. Conard *et al.*[1] (page 737 of this issue) now reveal that the Aurignacian inhabitants of the Swabian Jura had also mastered the art of music. Their detailed report highlights the discovery of a largely complete flute (Fig. 1) and two small flute fragments in the oldest Aurignacian layer at Hohle Fels Cave.

Conard recently reported[2] the discovery of a female figurine carved from mammoth ivory in an Aurignacian layer at Hohle Fels dated to at least 35,000 years ago (based on the newly calibrated radiocarbon timescale). At present, this is the earliest such find in the world. Additional examples of figurative art — of mammoths, horses, bison, cave lions, waterfowl and half-human, half-animal 'therianthropes' — have also been found in Aurignacian layers at Hohle Fels and other sites in the Swabian Jura. These finds suggest that the region was inhabited by a population of *Homo sapiens sapiens* that had mastered, among other things, the manipulation of mammoth ivory into three-dimensional, naturalistic forms for purposes not directly related to daily economic needs. Just as we continue to do today, these

H. JENSEN/UNIV. TÜBINGEN

Figure 1 | Sounds old. Conard *et al.*[1] have discovered the oldest known flute, at Hohle Fels Cave in Germany. The flute is made from bird bone, and dates from the early Aurignacian, 40,000 years ago.

© 2009 Macmillan Publishers Limited. All rights reserved

hunter–gatherers produced symbolic objects that embodied complex beliefs shared by a larger community of individuals.

The newly discovered flutes[1] suggest that music accompanied both daily and ritual activities. The most complete specimen, measuring 21.8 centimetres in length and with a diameter of 0.8 centimetres (Fig. 1), was produced from the radius (lower forelimb) of a griffon vulture. This flute retains five finger holes — although there may have been more — and the proximal end of the radius has been modified to serve as a mouthpiece. The two smaller fragments, made of ivory, are clearly derived from at least one other flute. There is little doubt that these implements are flutes, and given that they were recovered from secure, meticulously excavated and documented contexts within the cave, their archaeological association, stratigraphic provenance and age are not in question.

The oldest Aurignacian layer from which the three flute fragments were recovered dates to approximately 40,000 years ago and directly overlies the final Neanderthal layer. This date is believed to mark the initial expansion of modern human populations into the Swabian Jura, probably via the Danube Corridor[3]. although these are currently the earliest flutes known, it is reasonable to expect that even earlier examples were produced within and outside the region: the instruments from Hohle Fels are too 'evolved' in terms of design and manufacture to represent the first flutes. The makers and players of the Aurignacian flutes were thus not novices, but had considerable musical knowledge and experience that may have resulted from some form of transgenerational communication. Moreover, the earliest musical instruments, such as drums and rattles, were probably made of perishable materials — perhaps wood and hide — that are not routinely preserved in the archaeological record. Even so, these flutes from southwestern Germany are of immense importance, as they document a mature musical tradition that was firmly in place thousands of years earlier than previously thought.

The discoveries reported by Conard and colleagues[1] answer several crucial questions about the context and antiquity of early music in the Upper Palaeolithic. But precisely how and why music became such a ubiquitous — and economically profitable — aspect of virtually every modern human society is undear. Unlike the origins of language, which have long been the subject of intense research, the evolutionary significance of music has only recently been seriously investigated. Specifically, researchers seek to understand whether the human faculty for music is subject to natural selection. If so, when and under what circumstances did it evolve, and how might it have affected the reproductive fitness of individuals and groups that expressed musical behaviours?

Several general theories have been proposed to explain the evolution of music. For example, music is thought to have aided group cooperation, social cohesion and group synchrony, and coalition signalling[4–6]. It may also have played a part in mate selection, conflict reduction or vocal grooming[4]. Music could even have acted as a mnemonic device for long-term information exchange. But efforts to develop testable evolutionary hypotheses of music have been largely unsuccessful, and it is widely accepted that, if music is an evolutionary adaptation, then it probably had a complex origin that might be related to pre-existing cognitive and auditory adaptations in humans. Comparative research on the musical capacities of non-human animals[7] will allow us to develop a better understanding of which aspects of a general musical faculty, if any, are unique to humans.

The discovery of the flutes[1] from Hohle Fels make it clear that, by the early Aurignacian period roughly 40,000 years ago, our modern human predecessors in the Swabian Jura (and probably elsewhere) had thoroughly integrated music into their everyday lives — most probably as a critical element in rituals, but also as a means of fostering a sense of shared identity and common purpose. Music almost certainly helped to build and maintain group cohesion and social networks, by creating shared norms of musical aesthetics and storytelling, and through the strong emotions that music can elicit. Although we will never know precisely what music these Palaeolithic flautists played, or under what conditions they played it, Conard and colleagues' extraordinary finds[1] are clear proof that our ongoing obsession with music and musicians is of considerable antiquity. ■

Daniel S. Adler is in the Department of Anthropology, University of Connecticut, Storrs, Connecticut 06269, USA.
e-mail: daniel.adler@uconn.edu

1. Conard, N. J., Malina, M. & Münzel, S. C. *Nature* **460,** 737–740 (2009).
2. Conard, N. J. *Nature* **459,** 248–252 (2009).
3. Conard, N. J. & Bolus, M. *J. Hum. Evol.* **44,** 331–371 (2003).
4. Huron, D. *Ann. NY Acad. Sci.* **930,** 43–61 (2001).
5. Hauser, M. & McDermott, J. *Nature Neurosci.* **6,** 663–668 (2003).
6. Wiltermuth, S. S. & Heath, C. *Psychol. Sci.* **20,** 1–5 (2009).
7. Hagen, E. H. & Bryant, G. A. *Hum. Nature* **14,** 21–51 (2003).

STRUCTURAL BIOLOGY

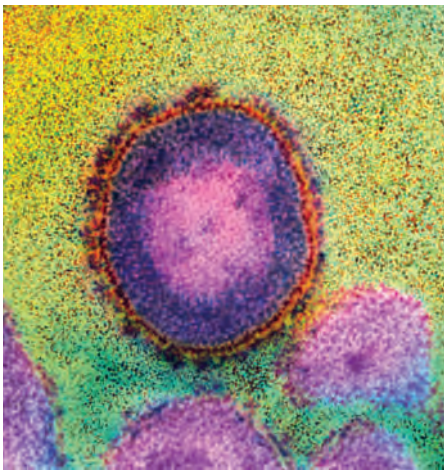
Aerial view of the HIV genome

Hashim M. Al-Hashimi

A bird's-eye view of the higher-order structure of HIV-1's entire RNA genome reveals new motifs in surprising places. Structural biologists can now zoom in on these regions to explore their functions further.

The genome of RNA viruses, such as the human immunodeficiency virus (HIV; Fig. 1), folds to form higher-order structures with stems and loops that contain motifs directing various steps of viral replication. Structural biologists usually 'cut out' these motifs and zoom in to determine their three-dimensional structures in an attempt to further understand their function. On page 711 of this issue, however, Watts *et al.*[1] zoom out and provide an 'aerial view' of the secondary structure of the entire HIV-1 genome. Using an innovative technique, they identify functional RNA motifs in surprising regions and define principles that govern the organization of the structure of the HIV-1 genome.

The methods commonly used to obtain detailed atomic-resolution images of biomolecules — nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography — have limitations that preclude analysis of the structure of entire RNA genomes. Both techniques rely on measuring the interactions between light and matter. In NMR spectroscopy, a solution containing the RNA of interest is immersed in a magnetic field and radio-frequencies are used to excite signals from its individual nuclei. As the size of the RNA under analysis increases, the signals become weaker and more congested, limiting structure determination of RNAs that are hundreds of nucleotides long.



NIBSC/SPL

Figure 1 | HIV. Transmission electron micrograph of a section through HIV. The virus is surrounded by an outer coat (red), and the RNA genome is enclosed in an inner protein core (pink).

X-ray crystallography measures the diffraction of X-rays when they strike crystals

© 2009 Macmillan Publishers Limited. All rights reserved

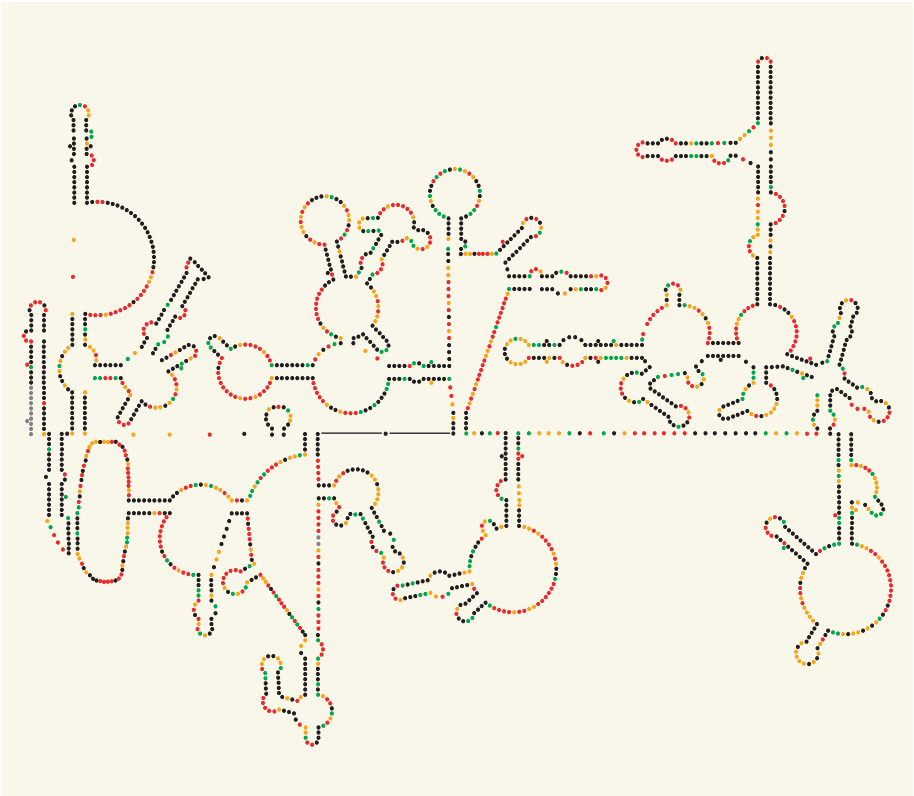


Figure 2 | The HIV-1 RNA genome shapes up. Secondary structure of a section of the HIV-1 RNA genome as determined by Watts *et al.*[1] using the SHAPE technique. Nucleotides are represented as coloured dots, with the colours depicting the amount of SHAPE reactivity, which reflects nucleotide flexibility and base pairing. The authors show that the genome has structured motifs (including stems and loops) in regulatory and protein-coding regions. Structured RNA in protein-coding regions may have a role in protein translation and in ensuring correct protein folding. SHAPE analysis of the entire HIV-1 genome is shown in Figure 2 on page 713.

containing ordered arrays of RNA. Although the technique does not suffer from size limitations, obtaining ordered crystals for highly flexible and diverse genomic RNA structures is a daunting task. High-resolution structure determination is also a time-consuming process, and as such is reserved for those privileged motifs that have been deemed functionally important. Consequently, more than 80% of the HIV-1 genome remains structurally uncharacterized.

Using a technique called SHAPE (selective 2′-hydroxyl acylation analysed by primer extension), Watts *et al.*[1] provide insight into the complete HIV-1 genome structure. The images produced are of lower resolution than those obtained by NMR spectroscopy and X-ray crystallography, but they span a much larger area of the genome. The technique is thus akin to zooming out on a map and getting a broader view of the landscape at the expense of fine details (Fig. 2).

SHAPE exploits the fact that the folds and loops of RNA molecules are stabilized by nucleotide base pairing. The technique relies on interactions between RNA and chemical reagents that react selectively, but sparsely, with flexible (unpaired) nucleotides, and which thereby modify their chemical structures. The reverse transcriptase enzyme is then used to make linear DNA copies of the RNA, which are terminated whenever the enzyme encounters a chemically modified site. By comparing the length and quantity of the DNA copies produced from RNA that has been exposed to chemical reagents with those from RNA that has not, the absolute reactivity, and thus flexibility, of individual nucleotides can be determined. Regions with low reactivity and flexibility correspond to regions of RNA with significant base-paired secondary structure, whereas regions with high reactivity and flexibility correspond to unpaired nucleotides.

Although there are various well-established chemical and enzymatic approaches for probing the secondary structure of RNA[2], many of the reagents used react with only a subset of nucleotides (for example, guanine and cytosine versus adenine and uracil) and their reactivity can depend on both RNA secondary structure and the intrinsic nucleotide activity. The unique reagents and chemistry used by SHAPE can help to surmount many of these limitations, and its automation allows for high-throughput studies of very large RNA structures. By plugging the SHAPE reactivity data into a computer algorithm that calculates the thermodynamic stability, and therefore the folded state, of RNA[3,4], Watts *et al.* propose a model of the secondary structure of the

50 YEARS AGO

Sir John Cockcroft is reported to have expressed the opinion on April 26 that in 1966 some 25 per cent of the requirements of the United Kingdom for electricity would be met by nuclear generation, 50 per cent by 1975 and 100 per cent by the end of the century. Questions asked in the House of Commons on June 8 indicate a disposition to allow political and social considerations to over-ride, if not distort, the technical and economic aspects, and there have been other attempts to make the effect on the coal industry the deciding factor in determining the development of nuclear power. The implications of technological change have been ignored, as has the effect of development on the cost of electricity supplied by nuclear power-stations.
From *Nature* 8 August 1959.

100 YEARS AGO

A survey of the progress made during the last twenty-five years in almost any field of engineering work would show an immense advance. Even during the past ten years very considerable progress has been made in certain branches of applied science, and in none of them to a greater extent than in the internal-combustion engine. We need not in this comparison claim the gun as a form of internal-combustion engine, though we are naturally entitled to do so. We may leave lethal weapons aside, and think only of the remarkable development of the reciprocating internal-combustion engine, and of the many changes it has brought about in our times. It has revolutionised cross-country transit. It has given us the long-deferred ... victory called the "conquest of the air." It is extraordinary to think of the numbers of men who have spent ingenious years in seeking a solution of the problem of flight. The solution has come in the unexpected form of a pair of long, sail-like arms, driven forward by a small high-speed internal-combustion engine.
From *Nature* 5 August 1909.

50 & 100 YEARS AGO

© 2009 Macmillan Publishers Limited. All rights reserved

entire HIV-1 RNA genome, which contains a dazzling 9,173 nucleotides.

The structured regions of the HIV-1 genome are concentrated in about 21 large domains[1]. Most of the functionally important structured RNA motifs that have been characterized so far reside in the untranslated (non-protein coding) region near the ends of the viral genome, which regulates viral replication and packaging of viral particles. Watts and colleagues[1] detect these previously characterized RNA motifs, sometimes as components of larger motifs, but they also identify structured RNA elements in protein-coding regions of the genome.

Many HIV-1 proteins are translated into polyprotein precursors by the ribosome as the viral RNA passes through it. The proteins are joined like beads on a string by linker peptides: these are later cleaved to release the individual proteins. There are also unstructured linker peptides between domains that make up the individual HIV proteins. Intriguingly, many of the newly identified structured RNA elements are located in regions that code for these flexible linkers. The authors propose that the structured RNAs slow down protein translation because these regions must be unfolded prior to entry into the ribosome. Because HIV proteins might be folded during translation — a process referred to as co-translational protein folding — this ribosomal pausing may provide additional time for proteins to adopt their correct three-dimensional structures.

This fascinating relationship between RNA and protein structure is not without precedent. The correlation between the secondary structure of messenger RNA and protein translation was recognized as early as three decades ago, and there are several studies showing that mRNA secondary structure can promote ribosomal pausing and modulate other aspects of translation and protein folding[5]. Watts and colleagues[1] go on to identify several other pause sites that seem to buy time at strategic moments during translation. For instance, unwinding folded RNA may allow binding of the signal-recognition particle to the elongating peptide chain. This protein–RNA particle guides the ribosome–peptide-chain complex to the endoplasmic reticulum for further processing. Ribosomal pausing may also provide time for frameshifting — whereby the ribosome stalls and skips over nucleotides without translating them, changing the reading frame — which allows translation of alternative HIV proteins from the same RNA.

On the other hand, highly unstructured regions are observed in hypervariable regions of the HIV-1 genome, which have important roles in viral host evasion. These unstructured regions are bordered by conserved and stable RNA structures that may help to prevent their interaction with the less variable neighbouring regions.

Whenever possible, Watts *et al.*[1] interpreted the details of the SHAPE model in relation to other biochemical and structural data, and information about evolutionary conservation of the pairing possibility of nucleotide regions. But there are still potential sources of error, particularly for an RNA structure of this size. Many regions probably do not exist as a single secondary structure, instead alternating between different conformations. The SHAPE-directed folding algorithm also fails to recognize some RNA structures, such as pseudoknots, or base pairs that form only as part of higher-order tertiary interactions. Atypical RNA structures may also interfere with the SHAPE analysis.

Notwithstanding these limitations, the study by Watts *et al.*[1] is a considerable achievement, showing the feasibility of obtaining 'aerial' views of large genomic RNA structures that reveal their architecture and possible functions. Structural biologists can now use this genomic map to judiciously zoom in on pieces of the HIV-1 genome and determine architectural and functional principles at the atomic level. Bridging these disparate RNA structure–function scales as well as moving towards movies of the genome in functional motion will be challenges for the future. But for now, it seems that the quest for a high-resolution structure of the entire HIV-1 RNA genome has begun in earnest. ■

Hashim M. Al-Hashimi is in the Department of Chemistry and Biophysics, University of Michigan, Ann Arbor, Michigan 48109-1055, USA.
e-mail: hashimi@umich.edu

1. Watts, J. M. *et al. Nature* **460,** 711–716 (2009).
2. Ehresmann, C. *et al. Nucleic Acids Res.* **15,** 9109–9128 (1987).
3. Mathews, D. H. *et al. Proc. Natl Acad. Sci. USA* **101,** 7287–7292 (2004).
4. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. *Proc. Natl Acad. Sci. USA* **106,** 97–102 (2009).
5. Kozak, M. *Gene* **361,** 13–37 (2005).

BIOGEOCHEMISTRY

Carbonate rocks deconstructed

Michael A. Arthur

The ratios of stable isotopes, especially isotopes of carbon and oxygen, have tales to tell about Earth's history. Post-depositional alteration of the carbonate rocks being studied may radically alter the story.

In a paper reminiscent of a scene from the movie *Total Recall*, in which Arnold Schwarzenegger rapidly terraforms Mars, Knauth and Kennedy (page 728 of this issue)[1] suggest that the 'greening' of Earth started about 850 million years ago in coastal regions. Their thinking is that the spread of photosynthetic life to the land then altered the chemical breakdown of rocks on Earth's surface[2], increasing the nutrient flux from land to ocean, and resulting in greater sequestration of organic matter in soils and marine sediments. A consequence of that, a hypothesized increase in atmospheric oxygen[3], would have set the stage for a subsequent and prolific increase in animal diversity.

These are intriguing ideas. Not least, Knauth and Kennedy provide a plausible explanation for the popular notion that oxygen levels increased substantially in the Neoproterozoic[3], the era that runs roughly from 1,000 million to 570 million years ago. But their conclusions will be controversial for at least two reasons.

First, Knauth and Kennedy present no direct evidence for the existence of a widespread terrestrial flora in the Neoproterozoic. Rather, in a creative step, they indirectly infer the occurrence of land colonization by photosynthetic organisms from changing geochemical patterns, as seen in stable isotopes of carbon, in Neoproterozoic shallow-marine carbonate rocks. Second, the authors argue that such rocks have nearly all experienced significant post-depositional alteration. In doing so, they cast doubt on the many studies that have aimed to reconstruct environmental conditions from the geochemical signals thought to have been locked in at the time that these carbonates were deposited. Their study even brings into question elements of the popular 'snowball Earth' hypothesis[4], which depends, in part, on the evidence for extreme variations in stable carbon isotopes from some of these same rocks.

So what exactly have Knauth and Kennedy done? Working from published sources, they have generated crossplots of the values of $\delta^{13}C$ (a measure of the ratio of ^{13}C to ^{12}C) and $\delta^{18}O$ (a measure of the ratio of ^{18}O to ^{16}O) in shallow-water limestones. They suggest that there is little difference between the distribution of values for much of the Neoproterozoic and the ensuing Phanerozoic, the interval that runs from 570 million years ago to the present. Thus, they argue, the post-depositional alteration (diagenesis) pathways and conditions for modern, Phanerozoic and most Neoproterozoic shallow-marine carbonates were the same.

Knauth and Kennedy[1] point out that, because of periodic falls in sea level, most recent and Phanerozoic shallow-marine carbonate rocks experienced extensive alteration through the action of fresh water. Fresh water percolates through such limestones, either in the near-surface unsaturated zone, or as groundwater 'lenses' that mix with sea water at their

© 2009 Macmillan Publishers Limited. All rights reserved

periphery. These low-salinity fluids drive dissolution of the more-soluble primary carbonates (typically aragonite and high-magnesium calcite today) and reprecipitation of less-soluble, low-magnesium calcite as cements and/or replacement minerals. These secondary calcites are lower in $\delta^{18}O$ because fresh water is depleted in ^{18}O relative to sea water. They are also lower in $\delta^{13}C$ because of the addition of soil-derived carbon dioxide, resulting from degradation (oxidation) of plant-derived organic matter that is highly depleted in ^{13}C, to the diagenetic fluids. With increasing addition of secondary cements, the bulk-rock, stable-isotope values decrease along a 'lithification' trend.

The authors[1] hypothesize that the paucity of $\delta^{13}C$ values lower than 0 parts per thousand (‰) in samples older than 850 million years, and the strong similarity between $\delta^{13}C$ distributions for carbonate rocks after this time in both the Neoproterozoic and Phanerozoic, indicate that a terrestrial biosphere was established by about 850 million years ago. In other words, the lower $\delta^{13}C$ values of late Neoproterozoic carbonate rocks, which are similar to those in the Phanerozoic, constitute evidence that there was a widespread photosynthetic biota, in coastal regions at least.

Indeed, it has long been appreciated that many shallow-water carbonate sediments experience changes that produce pronounced decreases in both the carbon and oxygen isotope values of the bulk rock[5,6]. Likewise, in reconstructing diagenetic history, carbonate sedimentologists have commonly used $\delta^{13}C$–$\delta^{18}O$ crossplots in studies of ancient carbonate rocks[7,8], including strata from the Proterozoic (which began 2,500 million years ago)[9]. Many such studies have used more than just stable isotopes to track diagenesis, and hint at complications. For example, the presence of concentrations of trace elements such as strontium, manganese and iron, as well as cathodoluminescence microscopy, have been used to eliminate heavily altered samples from consideration in interpreting primary seawater chemistry. Manganese and iron substitute into secondary carbonates because of the reducing conditions that characterize many subsurface diagenetic fluids. However, their incorporation in primary precipitates may also have occurred during the Proterozoic because of widespread deeper anoxic conditions and low sulphate concentrations[3]. Thus, the normal signals for carbonate diagenesis may not be entirely applicable to these older rocks.

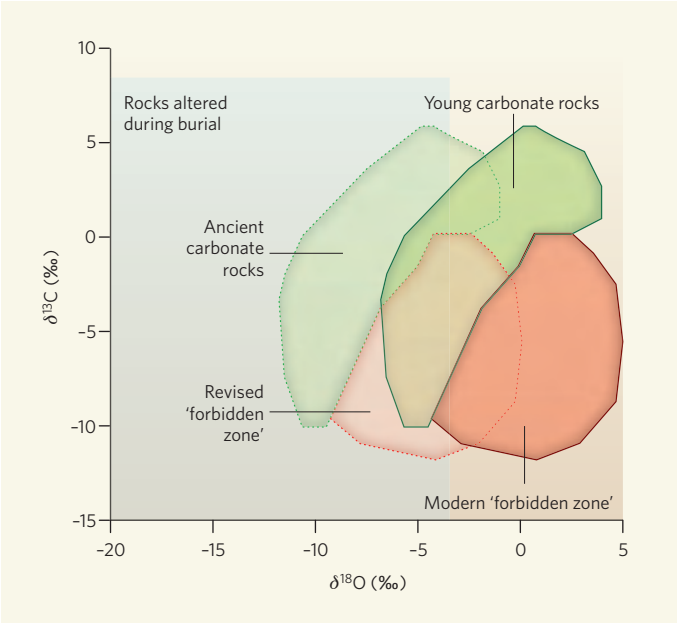


Figure 1 | Crossplot fields of $\delta^{13}C$ and $\delta^{18}O$ for shallow-marine carbonate rocks, and the 'forbidden zone'. Dark green, crossplot fields for young — modern — and many Phanerozoic rocks that have undergone alteration by 'freshwater' diagenesis soon after deposition. Light green, postulated field for ancient — some Phanerozoic and most Neoproterozoic — rocks that have experienced 'freshwater' diagenesis but were deposited with initially lower $\delta^{18}O$ values relative to modern rocks. The blue shaded zone delimits the crossplot area in which the original isotope values in the rock are thought to have become diagenetically altered (overprinted) during burial. Knauth and Kennedy[1] suggest that ancient carbonate rocks with extremely low carbon isotope values, which might indicate lack of overprinting alteration, should fall in a 'forbidden zone' (dark red). Few samples of ancient carbonate rocks fall in this zone, as defined by modern seawater conditions. But a challenge to the authors' assumption arises from the case that, in the past, combinations of lower average oxygen isotope values in sea water, and warmer temperatures, might plausibly have produced a $\delta^{18}O$ that was shifted 5‰ lower (light red). Many values for ancient carbonate rocks would then fall in this revised forbidden zone. Thus, they could represent relatively unaltered samples indicative of the carbon isotopic values for ancient sea water at the time.

Knauth and Kennedy[1], however, make certain assumptions about the primary values of $\delta^{18}O$ in all Neoproterozoic (and Phanerozoic) carbonates. In particular, they tacitly assume that the temperature and $\delta^{18}O$ of sea water in tropical to subtropical environments has not varied significantly through time. Thus, they expect that primary precipitates should have 'modern' values of $\delta^{18}O$ (actually, quite ^{18}O-enriched) (Fig. 1). This allows them to posit that the lack of Neoproterozoic samples that fall in their 'forbidden zone' (high $\delta^{18}O$ and low $\delta^{13}C$ values) indicates that the carbonate $\delta^{13}C$ values as low as −10‰, which some workers interpret as major perturbations of the $\delta^{13}C$ of seawater-dissolved inorganic carbon, possibly reflect a high degree of alteration.

Moving the 'modern' marine-lithification trend towards lower $\delta^{18}O$ values (Fig. 1), however, could change that interpretation. For example, if seawater $\delta^{18}O$ was somewhat lower or low-latitude temperatures warmer, primary carbonate $\delta^{18}O$ values in the warmest environments could have been in the −5 to −8‰ range. There are cogent arguments for[10] and against[11] significant variations through time in seawater $\delta^{18}O$, but there is little consensus in that regard.

Interpretation of stable isotopic values for late Neoproterozoic carbonate rocks is certainly a challenge. But there are other plausible models to account for the extreme carbon and oxygen isotope values that are seen in some strata, and that seem to represent global patterns[4]. For example, one proposed explanation for extreme negative carbon-isotope perturbations in the Neoproterozoic, representing a significant decrease in the carbon-isotopic difference between carbonate carbon and organic carbon, is carbon transfer from a massive pool of dissolved organic carbon from anoxic deep waters to the dissolved inorganic carbon pool of the oceans[12]. Another hypothesis to account for some of the same data is that enhanced oxidation of exposed organic-carbon-rich sediments on land resulted from a rise in atmospheric oxygen, with this ^{13}C-depleted dissolved inorganic carbon being incorporated into diagenetic carbonate cements in shallow-water limestones[13].

Knauth and Kennedy's work[1] will prompt much discussion, and it will be some time before it becomes evident how valid their ideas are. Meanwhile, there is a message for researchers studying carbonate rocks from the Proterozoic: they may do well to view the diagenetic history of such rocks in light of the principles applied to carbonate rocks from the Phanerozoic. ■

Michael A. Arthur is in the Department of Geosciences and the Penn State Astrobiology Research Center, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: arthur@geosc.psu.edu

1. Knauth, L. P. & Kennedy, M. J. *Nature* **460,** 728–732 (2009).
2. Kennedy, M., Droser, M., Mayer, L. M., Pevear, D. & Mrofka, D. *Science* **311,** 1446–1449 (2006).
3. Canfield, D. *Annu. Rev. Earth Planet. Sci.* **33,** 1–36 (2005).
4. Hoffman, P. F. & Schrag, D. P. *Terra Nova* **14,** 129–155 (2002).
5. Allan, J. R. & Matthews, R. K. *Sedimentology* **29,** 797–817 (1982).
6. Land, L. S. *US Geol. Surv. Bull.* **1578,** 129–137 (1986).
7. Lohmann, K. C. in *Paleokarst* (eds James, N. P. & Choquette, P. W.) 58–80 (Springer, 1988).
8. Brand, U. & Veizer, J. *J. Sedim. Petrol.* **51,** 987–997 (1981).
9. Jacobsen, S. B. & Kaufman, A. J. *Chem. Geol.* **161,** 37–57 (1999).
10. Veizer, J. *et al. Chem. Geol.* **161,** 59–88 (1999).
11. Muehlenbachs, K. *Chem. Geol.* **145,** 263–273 (1998).
12. Rothman, D. H., Hayes, J. M. & Summons, R. E. *Proc. Natl Acad. Sci. USA* **100,** 8124–8129 (2003).
13. Kaufman, A. J., Corsetti, F. A. & Varni, M. A. *Chem. Geol.* **237,** 47–63 (2007).

© 2009 Macmillan Publishers Limited. All rights reserved

PROGRESS

Beyond the myth of the supernova-remnant origin of cosmic rays

Yousaf Butt[1]

The origin of Galactic cosmic-ray ions has remained an enigma for almost a century. Although it has generally been thought that they are accelerated in the shock waves associated with powerful supernova explosions—for which there have been recent claims of evidence—the mystery is far from resolved. In fact, we may be on the wrong track altogether in looking for isolated regions of cosmic-ray acceleration.

Somewhere out in space, cosmic rays are mysteriously being hurled to extreme energies. The fastest of them travel very close to the ultimate speed limit: the speed of light. These particles, mostly protons, but also other ions and electrons, permeate our Galaxy and rain down on earth continuously, night and day. Although cosmic rays were discovered almost a century ago, back in the balloon age, their origins remain unclear even now. Almost no effort has been spared in pursuing this long-standing mystery: satellites, rockets and balloons have been launched, and enormous detector arrays have been installed on the ground and even under mountains and seas. One remarkable detector array, called IceCube, is several times larger than the Eiffel Tower and is buried more than a kilometre beneath the clear Antarctic ice.

Cosmic rays are divided into two main classes according to their energy and probable acceleration sites: those below about 10^{18} eV in energy are called Galactic cosmic rays (GCRs); above that energy, they are referred to as extragalactic cosmic rays, although the exact demarcation energy remains somewhat vague and debatable, and there could be some overlap. By comparison, the most powerful man-made particle accelerator, the Large Hadron Collider near Geneva, will reach energies of $\sim 10^{13}$ eV, which is barely one-ten-millionth of the most energetic cosmic ray recorded.

Here I exclusively discuss the lower-energy, Galactic, variety of cosmic rays. These particles are thought to be accelerated gradually, over centuries and even millennia, in the shock waves created by powerful supernova explosions within the Galaxy. As far back as 1953, Shklovskii speculated that "it is possible that ionized interstellar atoms are accelerated in the moving magnetic fields connected with an expanding [supernova remnant] nebula"[1]. However, we still have no proof of this scenario.

As GCR ions carry the bulk of the GCR energy, the main challenge rests in unambiguously identifying the origins of the ion, as opposed to the electron, component. Currently, the most direct way to find GCR acceleration sites is to look for telltale sources of γ-rays coincident with suspected celestial source sites. If an object is a GCR accelerator then it ought to have an overdensity of freshly accelerated cosmic rays in its vicinity. This 'cloud' of energetic cosmic rays can interact with the ambient matter and radiation to produce energetic γ-rays that can be detected on the Earth. The quandary is that both cosmic-ray ions and cosmic-ray electrons at the source site can produce the γ-rays we see, and it is extremely difficult to determine which type of particle was responsible for generating them.

There is fierce debate regarding the origin of the γ-rays seen in the direction of a few isolated supernova remnants (SNRs): whereas some advocate that ions are the source[2], others point out that electrons cannot be ruled out[3]. Though certainly fascinating, the outcome of this debate will not solve the puzzle of the origin of GCRs. Even if we eventually find direct evidence that some isolated SNRs are accelerating ions, this will not automatically prove that such objects are the main sources of GCRs, in general.

In fact, there are already sufficiently serious flaws in the standard picture that cosmic-ray ions originate in isolated SNRs that it is now necessary to entertain alternative, more comprehensive and realistic, ideas. In my view, the real challenge is not just finding individual cosmic-ray acceleration sites—a given cosmic ray need not even have a unique discrete site as its origin—but, rather, determining the integrated cosmic-ray acceleration process. This probably involves the entire Galaxy and its extended halo, together with its ensemble of isolated, as well as overlapping, SNRs (called superbubbles), in what may be considered a single holistic acceleration 'site'[4].

The standard model

Shock acceleration in isolated SNRs. The mechanism believed to be responsible for accelerating charged particles in an individual SNR is diffusive shock acceleration (DSA): such particles repeatedly scatter off magnetic turbulence on both sides of an SNR shock front, gaining speed as a result of the difference in the plasma velocities on either side of the shock. The greater the velocity difference, the greater the energy gained by the particle per shock crossing, and the larger the magnetic field (and turbulence), the higher the particle crossing frequency.

In the past few years, the HESS telescope array in Namibia and the CANGAROO array in Australia have discovered extended teraelectronvolt (TeV, 10^{12} eV) γ-ray emission from at least four isolated shell-type SNRs: RX J1713.7-3946, Vela Junior (RX J0852.0-4622), RCW 86 and SN 1006. Perhaps it is not coincidental that all four are relatively young, less than ~2,000 yr old. In these SNRs, there is a close correlation between the morphology of non-thermal X-ray and TeV emissions that suggests a common origin for the fluxes, namely the electrons[3,5,6], but viable ion emission models also exist[2].

Dynamical evidence for acceleration in SNRs. Two main effects are expected if an SNR shell is accelerating cosmic-ray ions: the physical separation between the forward shock and the 'contact discontinuity' (or reverse shock) should be considerably reduced, and the temperature at the forward shock should be depressed. Indeed, in images of Tycho's SNR made by NASA's Chandra X-ray Observatory, the separation between the forward shock and the contact discontinuity is smaller than expected—unless a significant fraction of the explosion

[1]High Energy Astrophysics Division, Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

©2009 Macmillan Publishers Limited. All rights reserved

energy has gone into the acceleration of cosmic-ray ions[7,8]. These measurements, along with evidence that the forward-shock temperature in the young remnant 1E 0102.2-7219 is lower than that expected from the measured expansion velocity[9], are indirect 'dynamical' evidence for the acceleration of cosmic-rays ions by SNR shocks. However, although ions may be being accelerated and may take up a large fraction of the SNR mechanical energy in these remnants, such evidence does not necessarily mean that the ions are being accelerated to TeV energies and beyond. And even if they are, as may be the case for part of SNR RCW 86's shell[10], it does not follow that isolated SNRs are the main source of cosmic rays.

Evidence for acceleration in old SNRs. Intriguingly, direct spectral signatures of GCR acceleration may have recently been seen in a handful of older SNRs, such as IC 443[11,12], W28[13] and perhaps also W41[14] and the recently discovered G353.6-0.7[15]. Both the MAGIC (in the Canary Islands) and VERITAS (in Arizona) telescopes have observed TeV γ-ray emission in the direction of IC 443[11,12]. This may be the signature of locally accelerated ions interacting with an abutting molecular cloud[16,17] (Fig. 1). However, there is also an energetic pulsar wind nebula not too distant from the TeV source region in IC 443, and it could be the ions accelerated by the pulsar—rather than by the SNR shock wave—that are diffusing out and powering the TeV emission in the adjacent cloud[18]. The spectrum of IC 443 measured by the EGRET instrument on board NASA's Compton Gamma Ray Observatory also appears to show a 'pion-hump' feature at about 70 MeV (see fig. 4 of ref. 19) possibly indicating ion acceleration and interaction there, although the statistics of the 'detection' are marginal at best. It will be very interesting to see if the Italian Space Agency's AGILE satellite and NASA's Fermi Gamma-ray Space Telescope (previously called GLAST), which are both orbiting γ-ray observatories, confirm this feature with higher significance. Could it be that these senior citizens of the SNR population ($\sim 10^4$–10^5 yr old) have a larger role in accelerating GCR ions than do the youths?[20]

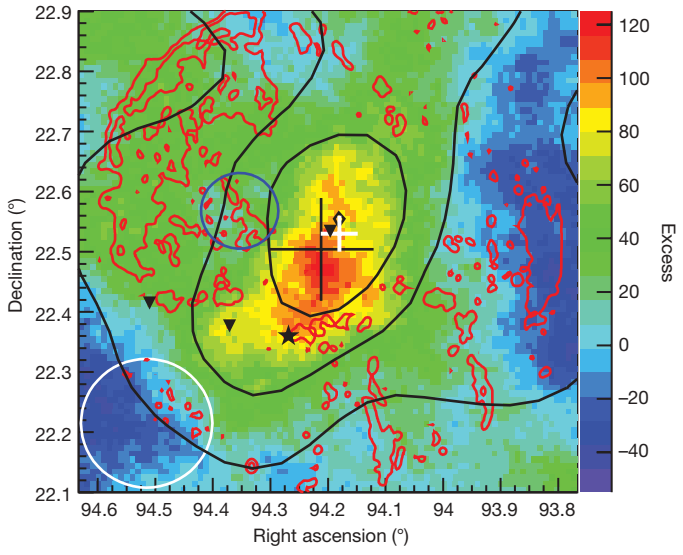


Figure 1 | Possible γ-ray emission from an SNR shock wave. An image, generated from VERITAS γ-ray telescope data, showing the very high-energy γ-ray signal detected in the direction of the SNR IC 443 (colour scale)[12]. The black cross indicates the centroid position of the extended TeV γ-ray emission, and its uncertainty. Similarly, the white cross indicates the position and uncertainty of the MAGIC telescope's TeV source[11] MAGIC J0616+225. The optical emission contours give an indication of the overall size of the SNR and are shown in red. The blue circle indicates the 95% error circle of the nearby lower-energy Fermi γ-ray source 0FGL J0617.4+2234. The intensity of the carbon monoxide emission, which is a proxy for the approximate amount of molecular cloud material in the vicinity of the SNR, is shown in black. The coincidence of TeV emission with the molecular material may be indicative of locally accelerated cosmic rays interacting with the ambient material. The locations of maser emission (possibly indicative of shock–cloud interactions) are shown by triangles, and the pulsar CXOU J061705.3+222127 is indicated by a star shape. The white circle illustrates the point spread function of the VERITAS telescope. For further details, see ref. 12. (Figure produced by B. Humensky for VERITAS[12]; reproduced by permission of the American Astronomical Society.)

Problems with the standard picture

The three strongest arguments supporting SNRs as possible cosmic-ray ion sources remain indirect: the theoretical spectrum of particles undergoing DSA, that is, a power-law spectrum with index of −2, supposedly agrees with that deduced from observations; SNRs are among the few Galactic sources that can satisfy the large energetic requirement of powering cosmic rays; and GCR electrons are observed to be accelerated in SNRs.

The first argument is not particularly compelling. Any accelerator in which a fractional gain in energy by some particles is accompanied by a fractional loss in the number of the remainder yields such a power-law spectrum[21]. This is a common feature of particles escaping from most acceleration regions, regardless of the precise processes involved. In any case, a joint analysis of the propagation and composition of cosmic rays favours[22] a source spectrum with a power-law index of around −2.35, not −2.0. More precise, nonlinear, versions of DSA make this discrepancy even worse, as they predict a 'concave' source spectrum with an index of about −1.5 at high energies[23]. Notably, DSA also faces significant empirical difficulties in explaining interplanetary particle acceleration[24,25].

Neither, to address the second argument, are SNRs the only source of mechanical energy in space: there is sufficient power, for instance, in Galactic rotation, which could also be tapped (perhaps by means of magnetic reconnection[21] or spiral density shocks[26]) to power GCRs. Other, more novel, sources of power include jets from accreting neutron stars and black holes, γ-ray bursts and pulsar outflows.

In any event, most of the SNR power injected into the Galaxy is not in the form of isolated SNRs, but rather in conglomerations of SNRs and massive stars (superbubbles), as outlined in the next section. If only individual SNRs were responsible for accelerating all cosmic rays, they would need to be extraordinarily efficient because they are so few. This distinction between SNRs and superbubbles would only be a minor issue if it were known that the putative process of cosmic-ray acceleration in superbubbles is the same as that thought to be at work in SNRs, but this remains far from certain.

As for the third point, the argument that SNRs are seen to be accelerating electrons does not automatically mean they are necessarily important sources of GCR ions. Because energetic ($\gtrsim$100-GeV) electrons lose energy much more rapidly than ions, sources of the GCR electrons seen on Earth are expected to be located, predominantly, within just a few kiloparsecs, whereas the acceleration regions of GCR ions are not similarly confined. (The entire Galaxy is about 30 kpc across.) GCR ions and electrons may have altogether different origins and acceleration processes.

The elemental and isotopic make-up of GCRs gives us further clues as to where they may originate. Such composition analysis indicates that they are accelerated mainly from a pool of old[27] ($\gtrsim 10^5$-yr-old), well-mixed interstellar material that does not reflect the elemental anomalies of fresh SNR ejecta[28]. In fact, isotopic anomalies in the GCR composition, such as the enhanced ^{22}Ne/^{20}Ne ratio, indicate[29] that GCRs are preferentially accelerated out of the material found in superbubbles.

The distribution of SNRs as a function of galactocentric distance also appears to be inconsistent with diffuse γ-ray data: the predicted cosmic-ray gradient, if due to SNRs alone, is steeper than that deduced from low-energy (<10-GeV) γ-ray maps. However, there may be ways to circumvent this problem[22]; for example, if more hydrogen gas exists farther from the centre of the Galaxy than is now thought, this could help redress this seeming inconsistency by compensating for the lower density of SNRs there.

©2009 Macmillan Publishers Limited. All rights reserved

Finally, the fact that GCRs are seen as coming from all directions with virtually equal intensity (that is, a 'low anisotropy', to better than 1 part in 1,000 at $\sim 10^{14}$ eV per nucleon) poses an even more serious challenge to the standard GCR origin picture. Such high-energy GCRs would be expected to escape from the Galaxy relatively quickly and, therefore, result in a greater anisotropy than has been observed, were they accelerated only by individual SNRs[30–32]. Isolated SNRs can be playing, at best, only a minor part in accelerating such cosmic-ray ions.

Thus both low-energy (<10-GeV) γ-ray data and high-energy (~100-TeV) cosmic-ray data argue against a supernova remnant origin of GCRs. Most of the power needed to accelerate GCRs may be supplied by SNRs, but the acceleration mechanism appears to be more complex and is probably distributed throughout the Galaxy and its extended halo[4,30–35].

Last year's intriguing report by the Milagro collaboration of a slight excess (a few parts in 10,000) of 10-TeV cosmic rays coming from two patches of the sky[36] does not detract from this conclusion. It remains to be seen whether the Milagro excess means that the solar system is being bathed in excess cosmic-ray nuclei from a nearby cosmic-ray accelerator[37], whether it is simply a local effect having to do with the sun's magnetic field structure[36] or whether it is something else entirely.

Distributed acceleration

Superbubbles. Most supernovae are of the core-collapse variety, having massive progenitor stars. Such massive stars are born in clusters of up to several thousand members, burn fast and die young, in near-simultaneous supernova explosions. The resulting multiple SNRs meld into one another, forming enormous superbubbles in the interstellar medium (ISM). Because most of the power injected by SNRs into the ISM is injected through such superbubbles, rather than through the isolated SNRs[38], it is imperative that we understand the role of superbubbles in GCR acceleration.

Unfortunately, the radiative signature of SNRs in the rarefied and hot medium of a superbubble interior is expected to be minimal[39], so we may be ignorant of most of the SNRs in the Galaxy. (These hidden SNRs might also explain the 'missing-SNR' problem, namely that we ought to see many more SNRs than have been detected so far.) Although our vantage point within the Galaxy makes it hard for us to observe and analyse the large and diffuse Galactic superbubbles, indirect evidence of GCR acceleration in superbubbles in the nearby Large Magellanic Cloud galaxy has recently been presented[40].

The current generation of TeV γ-ray telescopes could possibly detect some of the hidden SNRs in the Galaxy. Indeed, some of the many 'dark' extended TeV sources detected by the HEGRA and HESS collaborations could be such SNRs. The fact that the Milagro water-Cherenkov-detector group has detected extended γ-ray emission above 10 TeV coincident with some of these dark TeV sources makes them especially intriguing and worthy of pursuit as plausible GCR acceleration hotspots[41].

Large-scale putative GCR accelerators, such as superbubbles, would imply a spatially variable GCR intensity through the Galaxy, and the diffuse γ-ray maps made from data collected by the EGRET instrument do show possible evidence of this. Contiguous large-scale γ-ray features, uncorrelated with known Galactic molecular gas concentrations, have been detected and are significantly brighter than the average γ-ray background (ref. 42 and Supplementary Fig. 2 therein). Although these features have sometimes been interpreted in terms of a mysterious Galactic 'dark gas'[42] (not to be confused with dark matter or dark energy), an alternative explanation is that they may be simply a consequence of enhanced GCR source intensity there.

A striking example is the bright and extended excess-γ-ray region[42] coincident with the Gum nebula[43]. The Gum nebula is known to be internally powered by OB associations and/or SNRs[44]. A natural explanation for the excess diffuse γ-ray emission seen in this direction is that the cosmic-ray intensity there is significantly higher than the local one. Similarly, enhanced extended γ-ray emission is also coincident with a newly discovered superbubble in the constellation Ophiuchus[45]. It is most likely that the mysterious dark gas has different explanations in different regions of the sky: a spatially varying GCR intensity, recently detected molecular material not included in earlier models and perhaps some cold H I gas.

Galaxy-wide acceleration. Superbubbles cannot be the entire solution to the origin of GCRs, however, as they run into some of the same problems plaguing isolated SNRs. Superbubbles also cannot account for the low large-scale anisotropy of cosmic rays, nor explain the shallow cosmic-ray gradient deduced from γ-ray data. Even if isolated SNRs and superbubbles are considered the main power source for GCR acceleration, it is probable that the process is actually distributed across the Galaxy and the extended halo[30,31,33]. Cosmic-ray reacceleration may also be taking place at the Galactic-wind termination shock at a distance of a few hundred kiloparsecs from the centre of the Milky Way[34], as well as in 'slipping interaction regions' (50–100 kpc distant) in the Galactic wind[35]. The collective reacceleration of the cosmic-ray particles by this shock ensemble may also explain the observable cosmic-ray spectrum up to energies of $\sim 10^{17}$ eV, as well as the low anisotropy of high-energy GCRs[31].

Future prospects

The problem of the origin of cosmic rays is not that we have not yet found a firm spectral signature of ion acceleration in even a single isolated SNR, but that there are other, more severe, problems with this oversimplified scenario to begin with. Even if such a signature were found, it would not be sufficient to prove that isolated SNRs are the main accelerators of GCRs.

Because the process of GCR acceleration could be distributed[31,33], we may not even be posing the correct questions in trying to identify only discrete GCR source sites. A given cosmic ray need not have originated from exactly one source: its 'origin' may be intrinsically unclear.

What is needed is a better integration of 'microscopic' source-specific discrete acceleration models, with macroscopic Galactic—and extended halo, plus termination shock—propagation[22] and reacceleration[31] models, to provide a comprehensive picture of how GCRs gain energy. An early prototype of such an integrated model is the study in ref. 33. Although more sophisticated numerical models have since been developed (for example the popular GALPROP code[22]), important shortcomings remain. For instance, owing to the size of its numerical grid, GALPROP is currently unable to properly reproduce fine-scale spatial and temporal variations expected from localized sources of GCRs[46]. Much information about the origin of cosmic rays remains to be uncovered by modelling the Galactic 'ecology' of GCR acceleration, reacceleration, transport and composition holistically and at high fidelity.

A great deal of theoretical and observational work remains ahead. It would be useful, for instance, to understand whether the putative process of particle acceleration in superbubbles (that is, multiple interacting shocks embedded in pre-existing strong turbulence) could be as—or, perhaps, even more—efficient than that thought to operate in isolated SNR shocks. Supernovae provide the main energy source for superbubbles, but the details of the respective acceleration mechanisms in superbubbles and SNRs may be quite different. If, for example, the explosion energy is dumped into magnetic turbulence in the interior of a superbubble, it is conceivable that the superbubble, as a whole, acts as an accelerator with the second-order Fermi process (that is, stochastic acceleration) dominating, and with the potential for a large increase in maximum cosmic-ray energy because of the increase in spatial scale. On the other hand, if the energy stays in individual SNR blast waves, the regular DSA mechanism will dominate, with acceleration preferentially occurring across the superbubble wall. The resolution to the mystery of the origin of

©2009 Macmillan Publishers Limited. All rights reserved

GCRs will be incomplete until we have a better understanding of the role of superbubbles in GCR acceleration.

Observations of superbubbles will be equally important: for example, are any Galactic superbubbles γ-ray bright, indicating possible cosmic-ray acceleration there? Is there any evidence for non-thermal emission from them, as there is for several Large Magellanic Cloud superbubbles[40]? Are some of the extended hotspots attributed to the mysterious dark gas in diffuse γ-ray emission maps (ref. 42 and supplementary fig. 2 therein) really due to superbubbles? Are some of the dark TeV sources related to SNRs otherwise invisible because they are exploding within superbubbles[39]? Forthcoming maps of diffuse γ-ray emission recorded by the AGILE and Fermi observatories will go a long way towards answering some of these questions. Further in the future, neutrino observatories will also have an important role in understanding just how GCRs are accelerated.

Nevertheless, we should be prepared for the inevitable complications: for example, how do we find weak, diffuse GCR accelerators if the whole galaxy is aglow in γ-rays (or neutrinos), as it is? How important are discrete acceleration sites in comparison with distributed acceleration and reacceleration? How are we to discriminate between the γ-rays coming from extended acceleration regions and those arising from cosmic-ray propagation and interaction, in a Galaxy with a spatially variable cosmic-ray intensity? Is the usual assumption that the parameters of the cosmic-ray flux measured near Earth apply Galaxy-wide really correct? Unless we start asking the difficult questions, cosmic rays may hold fast to the secret of their origin for another century.

1. Shklovskii, I. S. On the origin of cosmic rays. *Dokl. Akad. Nauk SSSR* **91,** 475–478 (1953).
2. Berezhko, E. G. & Volk, H. Theory of cosmic ray production in the supernova remnant RX J1713.7-3946. *Astron. Astrophys.* **451,** 981–990 (2006).
3. Katz, B. & Waxman, E. In which shell-type SNRs should we look for gamma-rays and neutrinos from P–P collisions? *J. Cosmol. Astropart. Phys.* 01(2008)018 (2008).
4. Erlykin, A. D. & Wolfendale, A. W. The origin of cosmic rays. *J. Phys. G* **31,** 1475–1498 (2005).
5. Butt, Y. *et al.* X-ray hotspot flares and implications for cosmic ray acceleration and magnetic field amplification in supernova remnants. *Mon. Not. R. Astron. Soc.* **386,** L20–L22 (2008).
6. Liu, S. *et al.* Stochastic electron acceleration in shell-type supernova remnants. *Astrophys. J.* **683,** L163–L166 (2008).
7. Warren, J. *et al.* Cosmic-ray acceleration at the forward shock in Tycho's supernova remnant: evidence from Chandra X-ray observations. *Astrophys. J.* **634,** 376–389 (2005).
8. Völk, H. J., Berezhko, E. G. & Ksenofontov, L. T. Internal dynamics and particle acceleration in Tycho's SNR. *Astron. Astrophys.* **483,** 529–535 (2008).
9. Hughes, J. P., Rakowski, C. E. & Decourchelle, A. Electron heating and cosmic rays at a supernova shock from Chandra X-ray observations of 1E 0102.2-7219. *Astrophys. J.* **543,** L61–L65 (2000).
10. Helder, E. A. *et al.* Measuring the cosmic ray acceleration efficiency of a supernova remnant. *Science* doi:10.1126/science.1173383 (25 June 2009).
11. Albert, J. *et al.* Discovery of very high energy gamma radiation from IC 443 with the MAGIC telescope. *Astrophys. J.* **664,** L87–L90 (2007).
12. Acciari, V. A. *et al.* (VERITAS collaboration). Observation of extended vhe emission from the supernova remnant IC 443 with VERITAS, *Astrophys. J. Lett.* (in the press).
13. Aharonian, F. *et al.* Discovery of very high energy gamma-ray emission coincident with molecular clouds in the W 28 (G6.4–0.1) field. *Astron. Astrophys.* **481,** 401–410 (2008).
14. Tian, W. *et al.* VLA and XMM-Newton observations of the SNR W41/TeV gamma-ray source HESS J1834–087. *Astrophys. J.* **657,** L25–L28 (2007).
15. Tian, W. *et al.* Discovery of the radio and X-ray counterpart of TeV γ-ray source HESS J1731–347. *Astrophys. J.* **679,** L85–L88 (2008).
16. Torres, D. F. *et al.* MAGIC J0616+225 as delayed TeV emission of cosmic rays diffusing from the supernova remnant IC 443. *Mon. Not. R. Astron. Soc.* **387,** L59–L63 (2008).
17. Zhang, L. & Fang, J. Nonthermal emission from a radio-bright shell-type supernova remnant IC 443. *Astrophys. J.* **675,** L21–L24 (2008).
18. Bartko, H. & Bernarek, W. γ-ray emission from PWNe interacting with molecular clouds. *Mon. Not. R. Astron. Soc.* **385,** 1105–1109 (2008).
19. Gaisser, T. *et al.* Gamma-ray production in supernova remnants. *Astrophys. J.* **492,** 219–227 (1998).
20. Yamazaki, R. *et al.* TeV γ-rays from old supernova remnants. *Mon. Not. R. Astron. Soc.* **371,** 1975–1982 (2006).
21. Colgate, S. & Li, H. in *The Role of Neutrinos, Strings, Gravity, and Variable Cosmological Constant in Elementary Particle Physics* (eds Kursunoglu, B. N., Mintz, S. L. & Perlmutter, A.) 149–155 (Kluwer Academic, 2001).
This manuscript presents some speculative, although truly novel and original, ideas regarding cosmic-ray acceleration.
22. Strong, A. W., Moskalenko, I. V. & Ptuskin, V. S. Cosmic-ray propagation and interactions in the Galaxy. *Annu. Rev. Nucl. Part. Syst.* **57,** 285–327 (2007).
23. Ellison, D. C., Berezhko, E. G. & Baring, M. G. Nonlinear shock acceleration and photon emission in supernova remnants. *Astrophys. J.* **540,** 292–307 (2000).
24. Krimigis, S. M. Voyager energetic particle observations at interplanetary shocks and upstream of planetary bow shocks – 1977–1990. *Space Sci. Rev.* **59,** 167–201 (1992).
25. Fisk, L. A. & Gloeckler, G. Thermodynamic constraints on stochastic acceleration in compressional turbulence. *Proc. Natl Acad. Sci. USA* **104,** 5749–5754 (2007).
26. Duric, N. The origin of cosmic rays in spiral galaxies. *Space Sci. Rev.* **48,** 73–111 (1988).
27. Wiedenbeck, M. E. *et al.* Constraints on the time delay between nucleosynthesis and cosmic-ray acceleration from observations of 59Ni and 59Co. *Astrophys. J.* **523,** L61–L64 (1999).
28. Wiedenbeck, M. *et al.* in *Proc. 28th Int. Cosmic Ray Conf.* (eds Kajita, T., Asaoka, Y., Kawachi, A., Matsubara, Y. & Sasaki, M.) 1899–1902 (Universal Academy, 2003).
29. Binns, W. R. *et al.* OB associations, Wolf Rayet stars, and the origin of Galactic cosmic rays. *Space Sci. Rev.* **130,** 439–449 (2007).
This paper presents arguments for superbubbles having a significant role in Galactic cosmic-ray acceleration, on the basis of the composition of cosmic rays.
30. Parizot, E., Paul, J. & Bykov, A. M. in *Proc. 27th Int. Cosmic Ray Conf.* 2070–2073 (Copernicus Gesellschaft, 2001).
Many of the hidden assumptions that have perpetuated the myth that cosmic rays most likely originate in isolated supernova remnants are outlined in this work.
31. Seo, E. S. & Ptuskin, V. S. Stochastic reacceleration of cosmic rays in the interstellar medium. *Astrophys. J.* **431,** 705–714 (1994).
32. Hillas, M. J. Can diffusive shock acceleration in supernova remnants account for high-energy galactic cosmic rays? *J. Phys. G* **31,** R95–R131 (2005).
33. Medina-Tanco, G. A. & Opher, R. Spatial and temporal distributed acceleration of cosmic rays by supernova remnants three-dimensional simulations. *Astrophys. J.* **411,** 690–707 (1993).
This study suggests that distributed Galaxy-wide acceleration could be an acceptable mechanism for Galactic cosmic-ray acceleration.
34. Zirakashvili, V. N. & Völk, H. J. Cosmic ray reacceleration on the galactic wind termination shock. *Adv. Space Res.* **37,** 1923–1927 (2006).
35. Völk, H. J. & Zirakashvili, V. N. Cosmic ray reacceleration by spiral shocks in the galactic wind. *Astron. Astrophys.* **417,** 807–817 (2004).
36. Abdo, A. *et al.* (Milagro Collaboration). Discovery of localized regions of excess 10-TeV cosmic rays. *Phys. Rev. Lett.* **101,** 221101 (2008).
37. Drury, L. & Aharonian, F. The puzzling MILAGRO hot spots. *Astropart. Phys.* **29,** 420–423 (2008).
38. Higdon, J. C. & Ligenfelter, R. E. OB associations, supernova-generated superbubbles, and the source of cosmic rays. *Astrophys. J.* **628,** 738–749 (2005).
39. Tang, S. & Wang, Q. D. Supernova blast waves in low-density hot media: a mechanism for spatially distributed heating. *Astrophys. J.* **628,** 205–209 (2005).
40. Butt, Y. M. & Bykov, A. M. A cosmic-ray resolution to the superbubble energy crisis. *Astrophys. J.* **677,** L21–L22 (2008).
41. Abdo, A. A. *et al.* TeV gamma-ray sources from a survey of the galactic plane with Milagro. *Astrophys. J.* **664,** L91–L94 (2007).
42. Grenier, I. A., Casandjian, J.-M. & Terrier, R. Unveiling extensive clouds of dark gas in the solar neighborhood. *Science* **307,** 1292–1295 (2005).
43. Woermann, B. *et al.* Kinematics of the Gum nebula region. *Mon. Not. R. Astron. Soc.* **325,** 1213–1227 (2001).
44. Yamaguchi, N. *et al.* Distribution and kinematics of the molecular clouds in the Gum nebula. *Publ. Astron. Soc. Jpn.* **51,** 765–774 (1999).
45. Pidopryhora, Y., Lockman, F. J. & Shields, J. C. The Ophiuchus superbubble: a gigantic eruption from the inner disk of the Milky Way. *Astrophys. J.* **656,** 928–942 (2007).
46. Büsching, I. *et al.* Cosmic-ray propagation properties for an origin in supernova remnants. *Astrophys. J.* **619,** 314–326 (2005).

Acknowledgements Part of this work was carried out while the author was a fellow at the National Academy of Sciences. The support of a NASA Long Term Space Astrophysics grant is gratefully acknowledged.

Author Information Correspondence should be addressed to Y.B. (ybutt@cfa.harvard.edu).

©2009 Macmillan Publishers Limited. All rights reserved

ARTICLES

miR-145 and miR-143 regulate smooth muscle cell fate and plasticity

Kimberly R. Cordes[1,2,3], Neil T. Sheehy[1,2,3], Mark P. White[1,2,3], Emily C. Berry[1,2,3], Sarah U. Morton[1,2,3], Alecia N. Muth[1,2,3], Ting-Hein Lee[4], Joseph M. Miano[4], Kathryn N. Ivey[1,2,3] & Deepak Srivastava[1,2,3]

MicroRNAs (miRNAs) are regulators of myriad cellular events, but evidence for a single miRNA that can efficiently differentiate multipotent stem cells into a specific lineage or regulate direct reprogramming of cells into an alternative cell fate has been elusive. Here we show that miR-145 and miR-143 are co-transcribed in multipotent murine cardiac progenitors before becoming localized to smooth muscle cells, including neural crest stem-cell-derived vascular smooth muscle cells. miR-145 and miR-143 were direct transcriptional targets of serum response factor, myocardin and Nkx2-5 (NK2 transcription factor related, locus 5) and were downregulated in injured or atherosclerotic vessels containing proliferating, less differentiated smooth muscle cells. miR-145 was necessary for myocardin-induced reprogramming of adult fibroblasts into smooth muscle cells and sufficient to induce differentiation of multipotent neural crest stem cells into vascular smooth muscle. Furthermore, miR-145 and miR-143 cooperatively targeted a network of transcription factors, including Klf4 (Kruppel-like factor 4), myocardin and Elk-1 (ELK1, member of ETS oncogene family), to promote differentiation and repress proliferation of smooth muscle cells. These findings demonstrate that miR-145 can direct the smooth muscle fate and that miR-145 and miR-143 function to regulate the quiescent versus proliferative phenotype of smooth muscle cells.

MicroRNAs represent a class of small (~20–25 nucleotides), non-coding RNAs that are key regulators of many cellular events, including the balance between proliferation and differentiation during tumorigenesis and organ development[1–3]. MicroRNAs are initially transcribed as longer primary transcripts (pri-miRNAs) and processed first by the RNase enzyme complex, Drosha–DGCR8, and then by Dicer, leading to incorporation of a single strand into the RNA-induced silencing complex. Each of the ~650 human miRNAs is predicted to interact with more than 100 target mRNAs in a sequence-specific fashion involving Watson–Crick base-pairing among nucleotides 2–8 of the miRNA[4,5]. MicroRNAs generally inhibit target messenger RNAs by repressing translation or reducing mRNA stability. MicroRNAs may also activate mRNA translation under certain cellular conditions[6].

Regulation of cardiovascular cell fate decisions by miRNAs and control of proliferation and differentiation in cardiac progenitors have been reported, but remain inefficient[7–12]. A multipotent cardiac progenitor pool that can differentiate into cardiac myocytes, vascular smooth muscle cells (VSMCs) and endothelial cells exists[13], as do multipotent neural crest stem cells that can also give rise to VSMCs, as well as melanocytes, chondrocytes and neurons[14]. Among these cell types, VSMCs are uniquely plastic, as they can oscillate between a proliferative or a quiescent, more differentiated state[15]. This plasticity contributes to many human vascular diseases, including atherosclerosis[16,17]. The transcriptional control of this oscillation has been described[16], but whether VSMC-enriched miRNAs exist or participate in this process is unknown. Here we demonstrate that miR-145 and miR-143 are tightly integrated into a core transcriptional network involved in smooth muscle differentiation and proliferation, and that miR-145 functions as a critical switch in promoting smooth muscle differentiation.

miR-143 and miR-145 expression

We reported that miR-143 is the most enriched miRNA during differentiation of mouse embryonic stem (ES) cells into multipotent cardiac progenitors[11]. *miR-143* is highly conserved and lies within 1.7 kilobases (kb) of another conserved miRNA, *miR-145*, on mouse chromosome 18 (Supplementary Fig. 1a, b). Both miRNAs are downregulated in various cancer cell lines, colon cancers, and lung cancers[2]. Given their genomic organization and proximity, *miR-143* and *miR-145* may be contained in a bicistronic primary transcript, but we were unable to amplify a common transcript from RNA, possibly because pri-miRNA transcripts are rapidly processed into their mature forms. *DGCR8*-null ES cells lack nuclear miRNA processing activity and have a defect in differentiation[18], but can form mesoderm. Using primers for each miRNA and RNA from *DGCR8*-null embryoid bodies (EBs), we generated an amplicon that encompassed both miRNAs (Supplementary Fig. 1b), suggesting that *miR-143* and *miR-145* were transcribed as a bicistronic unit and therefore share common regulatory elements that control their expression.

To determine if these two miRNAs are also enriched in multipotent cardiac progenitors *in vivo*, we bred transgenic mice containing *Cre recombinase* in the *Islet1* locus[19] with *Rosa26-EYFP* mice[20], and isolated YFP^+ cardiac progenitor cells at embryonic day (E)9.5 by fluorescence-activated cell sorting (FACS) (Supplementary Fig. 1c, d). The *Islet1-Cre* mice mark early multipotent cardiac progenitor cells that can differentiate into cardiac muscle, smooth muscle and endothelial cells[21]. Quantitative RT–PCR (qPCR) revealed that miR-143 and miR-145 were enriched in YFP^+ cells (Supplementary Fig. 1e). qPCR with RNA from mouse hearts or whole embryos at varying stages of development also revealed enrichment of both miRNAs throughout cardiogenesis, before being downregulated in the adult heart (Supplementary Fig. 1f, g).

[1]Gladstone Institute of Cardiovascular Disease, San Francisco, California 94158, USA. [2]Department of Pediatrics, University of California, San Francisco, California 94543, USA. [3]Department of Biochemistry & Biophysics, University of California, San Francisco, California 94143, USA. [4]Aab Cardiovascular Research Institute, University of Rochester School of Medicine and Dentistry, Rochester, New York 14642, USA.

©2009 Macmillan Publishers Limited. All rights reserved

Transcriptional regulation of *miR-143/miR-145*

To identify the tissue-specific expression and regulation of *miR-143/miR-145* during mouse development, we searched for upstream regulatory regions. Comparison of genomic sequences across species revealed a 4.2-kb genomic region upstream of *miR-143/miR-145* that was highly conserved between human and mouse (Supplementary Fig. 2a) and directed LacZ reporter expression specifically in multipotent cardiac mesodermal progenitors of transgenic mice as early as E7.5 (Fig. 1a, b). By E9.5, LacZ expression was more robust and uniform in the heart and outflow tract and in cardiac progenitors of the pharyngeal mesoderm; expression was also present in the aorta just as smooth muscle differentiation began, but was absent in the cardinal vein (Fig. 1c, d). LacZ expression was robust in the endocardium and myocardium (Fig. 1d). During later cardiogenesis, expression became restricted to the ventricles and atria, but was notably absent in the aorta and pulmonary arteries (Fig. 1e). Postnatally, the pattern was reversed, with high transcript levels in smooth muscle of the aorta, pulmonary artery and coronary vessels but undetectable levels in the ventricular myocardium (Fig. 1f–h). This enhancer was also active in the smooth muscle of the intestines (Supplementary Fig. 2b, c). The enhancer recapitulated the endogenous miR-145 expression, with transcripts in the smooth muscle of the adult aorta and coronary artery, but not ventricular myocardium, as shown by section *in situ* hybridization (Fig. 1i, j and Supplementary Fig. 2d).

Deletions of the 4.2-kb *miR-143/miR-145* enhancer revealed that a 0.9-kb region was sufficient for miR-143/miR-145 cardiac and smooth muscle expression (Fig. 2a–e, Supplementary Fig. 3a). Within this regulatory region, we observed *cis* elements highly conserved between human, mouse and zebrafish that represented potential binding sites for the essential cardiac transcription factors, serum response factor (SRF) and Nkx2-5 (Supplementary Fig. 3b). SRF plays a dual role in cardiac and VSMC development, influencing both proliferation and differentiation depending on the types of co-activators or repressors present at specific developmental or cellular stages[22]. The potent SRF co-activator, myocardin (Myocd)[23], is a component of a molecular switch for the VSMC fate[24] and is sufficient to effect both structural and physiological attributes of this cell type[25]. SRF weakly activated the *miR-143/miR-145* enhancer upstream of a luciferase reporter, but co-transfection of Myocd synergistically and robustly activated luciferase activity in Cos cells (Fig. 2f). Mutation of the highly conserved CArG box in the SRF binding site decreased Myocd-dependent luciferase activity (Fig. 2f). Nkx2-5 could also independently activate this enhancer, and the combination of SRF, Myocd and Nkx2-5, which also interacts with SRF[26], had additive effects on luciferase activity. Mutation of each binding site progressively decreased luciferase activity (Fig. 2f).

In vivo, mutation of the SRF binding site disrupted lacZ expression in the outflow tract and aorta, while disruption of the Nkx2-5 binding site diminished expression in the ventricles and atria (Fig. 2c, d), suggesting modular regulation by the enhancer. Mutation of both the SRF and Nkx2-5 binding sites abolished all activity of the enhancer within the heart (Fig. 2e). VSMC and atrial expression postnatally was also dependent upon the SRF-binding *cis* element (Supplementary Fig. 3c). Electromobility shift assay confirmed that SRF could specifically bind to the putative binding site in the *miR-143/miR-145* enhancer (Supplementary Fig. 3d). Furthermore, miR-143 and miR-145 were each expressed at lower levels in *SRF*-null EBs compared to wild-type EBs derived from the respective ES cells (Fig. 2g). The levels were also reduced in mesoderm-rescued *SRF*-null EBs[11], confirming that the decreases did not reflect the absence of mesoderm (Fig. 2g). Similarly, miR-143 and miR-145 expression was also decreased in hearts of *Nkx2-5* mutant mouse embryos in a dose-dependent fashion (Fig. 2h).

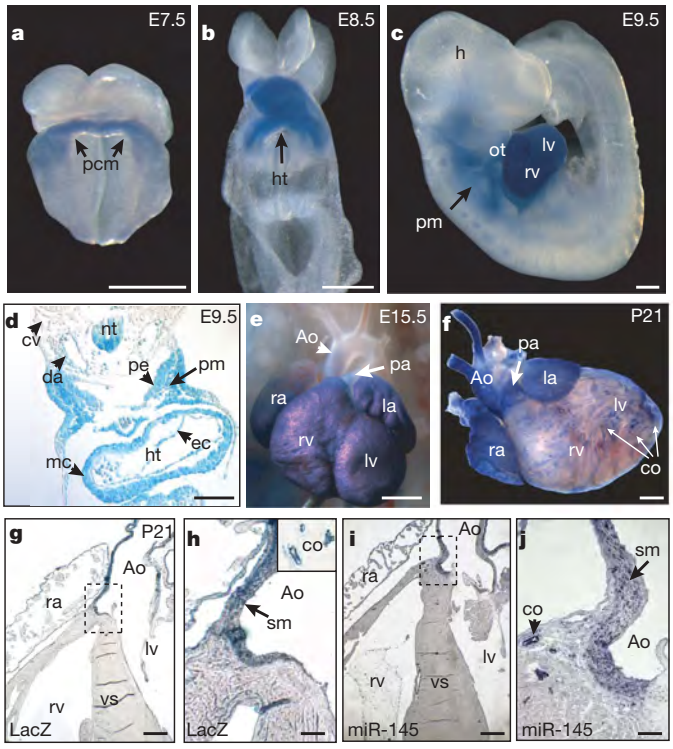


Figure 1 | miR-143 and miR-145 are cardiac-specific and smooth-muscle-specific miRNAs. a–c, Whole mounts showing cardiac-specific β-gal activity in transgenic mouse embryos with a 4.2-kb enhancer-lacZ construct (Supplementary Fig. 2a) at indicated time points. **d**, Transverse section of **c** showing β-gal expression in pharyngeal mesoderm (pm), pharyngeal endoderm (pe), dorsal aorta (da), myocardium (mc), endocardium (ec). **e**, **f**, β-gal expression in E15.5 (**e**) or post-natal day (P)21 (**f**) heart. Scale bars: 200 μm (**a**, **b**), 100 μm (**c–e**), 1 mm (**f**). Ao, aorta; pa, pulmonary artery. **g**, **h**, Transverse sections of **f**; co, coronary artery. **i**, **j**, Section in situ hybridization of miR-145 in P21 heart section. Scale bars: 500 μm (**g**, **i**), 100 μm (**h**, **j**). Inset in **h** highlights co. Panels **h** and **j** represent higher magnification of boxed areas. pcm, precardiac mesoderm; ht, heart; h, head; ot, outflow tract; rv, right ventricle; lv, left ventricle; cv., cardinal vein; ra, right atrium; la, left atrium; vs, ventricular septum.

Dysregulation in vascular disease

The dynamic stage-dependent expression of miR-143 and miR-145 raised the possibility that their expression may also vary with the oscillation of VSMCs between differentiated and proliferative phenotypes. In a mouse model of this proliferative switch, ligation of the carotid arteries typically results in narrowing of the vascular lumen as a result of phenotypic modulation and proliferation of VSMCs. qPCR revealed marked decreases in miR-143 and miR-145 expression in injured carotid arteries compared to contralateral control vessels (Fig. 2i). miR-21 expression was increased as expected[27], and miR-16 was unchanged, demonstrating the presence of intact small RNAs. *In situ* hybridization of injured and control carotid arteries also revealed marked downregulation of miR-143 and miR-145 expression in the thickened vascular wall, coincident with decreased expression of the differentiation marker, smooth muscle α-actin (Sm-α-actin) (Supplementary Fig. 4a). As a control, miR-143 and miR-145 levels were unchanged in cardiac muscle after injury (Supplementary Fig. 4b). Interestingly, transcripts of miR-145 were also downregulated to nearly undetectable levels in atherosclerotic lesions containing neointimal hyperplasia (Fig. 2j).

Regulation of cell fate and proliferation

The bimodal expression of miR-143 and miR-145 early during VSMC induction, and subsequently during the maturation into a non-proliferating, differentiated phenotype, led us to investigate their potential function in these settings. As miR-145 and miR-143 expression was directly activated by SRF-Myocd, we first investigated whether expression of either miRNA was necessary for Myocd-induced

©2009 Macmillan Publishers Limited. All rights reserved

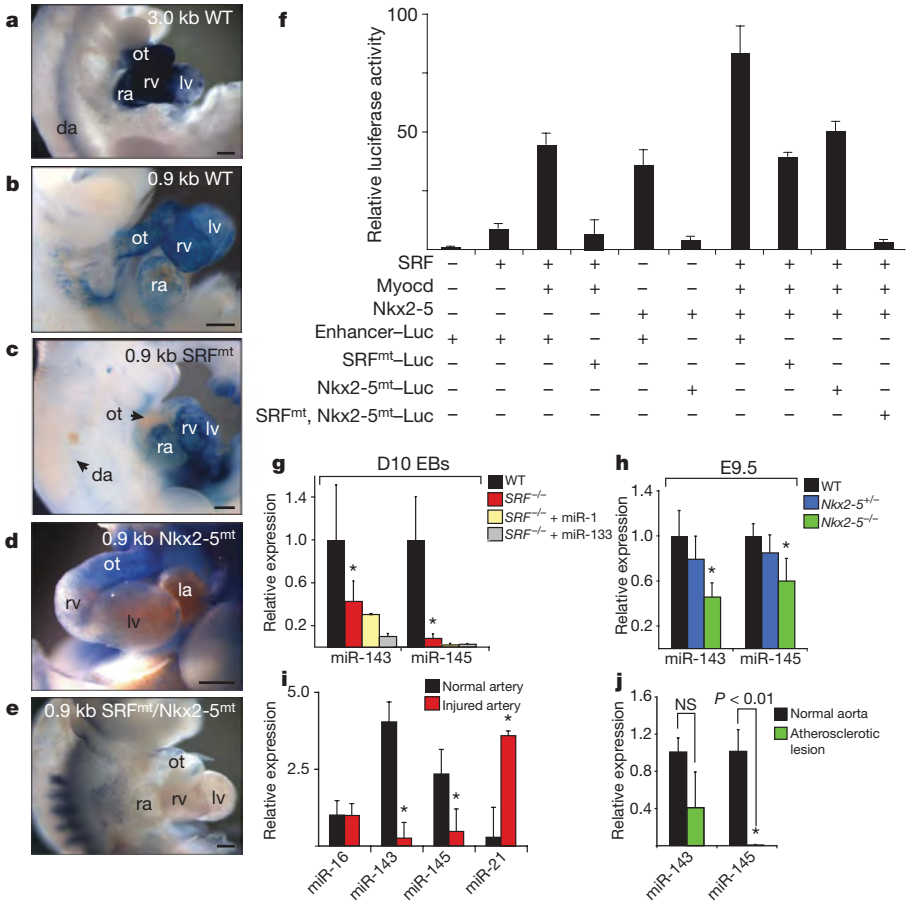


Figure 2 | SRF and Nkx2-5 directly regulate cardiac and smooth muscle expression of miR-143 and miR-145. **a**–**e**, Lateral (**a**–**c**, **e**) or frontal (**d**) cardiac views of transgenic embryos containing indicated wild type (WT) or mutant (mt) lacZ constructs and stained for β-Gal activity. **f**, Fold-activation of luciferase activity directed by introduction of SRF, Myocd or Nkx2-5 expression vectors with the miR-143/145 enhancer in Cos cells. All changes were statistically significant with $P < 0.05$ ($n = 5$). **g**, miR-143 and miR-145 expression levels assessed by qPCR on day 10 embryoid bodies (EBs) of indicated genotypes. **h**, qPCR of miR-143 and miR-145 in *Nkx2-5*$^{+/-}$ and *Nkx2-5*$^{-/-}$ E9.5 hearts relative to wild type. **i**, **j**, qPCR of miRNAs in injured vessels (**i**) or atherosclerotic lesions (**j**) compared to normal arterial expression. Results shown in (**g**–**j**) are from three experiments. ot, outflow tract; ra, right atrium; lv, left ventricle; rv, right ventricle; la, left atrium; da, dorsal aorta. NS, not significant; *$P < 0.05$; error bars, s.d. Scale bars: 50 μm.

reprogramming of fibroblasts into VSMCs. Introduction of 1–2 μg of Myocd into fibroblasts reliably resulted in >50% conversion to VSMCs[22]. Inhibition of miR-145 using cholesterol-modified antisense oligonucleotides (antagomirs)[28] blocked Myocd's ability to convert fibroblasts into VSMCs, as illustrated by Sm-α-actin immunostaining and expression of multiple smooth muscle markers assessed by qPCR and western blot (Fig. 3a–c, e). The knockdown of miR-143 had little effect on Myocd-induced smooth muscle conversion (Fig. 3a, b and Supplementary Fig. 5a, b). Neither miRNA was sufficient to reprogram fibroblast cells. However, miR-145 potentiated Myocd's reprogramming effects. Although 50 ng of Myocd was insufficient to induce VSMC gene expression, simultaneous addition of miR-145, but not miR-143, resulted in robust VSMC differentiation, equivalent to that observed with 1–2 μg of Myocd (Fig. 3b, d, e and Supplementary Fig. 5c). Thus, miR-145 activity was required for Myocd-dependent conversion of fibroblasts into VSMCs, and miR-145 robustly potentiated Myocd's effects.

To test an alternative cell type in which miR-145 may be sufficient for VSMC differentiation, we used a multipotent neural crest stem cell line that can differentiate into numerous cell types (for example, melanocytes, chondrocytes, neurons), including VSMCs, on exposure to 5 days of TGF-β[29]. Remarkably, introduction of miR-145, but not miR-143, into neural crest stem cells was sufficient to guide ~75% of cells into the VSMC lineage within only 24 hours, as determined by immunocytochemistry with multiple markers (Fig. 3f). qPCR and western blot revealed upregulation of numerous markers of VSMC differentiation, including Sm-α-actin, Sm-22α, and smooth muscle myosin heavy chain (Sm-MHC) by miR-145 but not miR-143 (Fig. 3g, h; Supplementary Fig. 5d, e). The neural crest stem-cell-derived VSMCs exhibited calcium flux measurements similar to cultured VSMCs in response to endothelin-1 stimulation, indicating the differentiation of functionally mature smooth muscle (Fig. 3i). Thus, miR-145 was sufficient for directing the VSMC fate from multipotent neural crest stem cells that normally populate the aortic smooth muscle tissue, where miR-145 is expressed.

miRNA targets and mechanism

The mechanism by which these miRNAs regulate VSMCs is dependent on their mRNA targets. A bioinformatics approach, incorporating sequence matching and mRNA secondary structure to predict mRNA targets (K.N.I. and D.S., unpublished results; also see Methods), revealed multiple highly conserved binding sites for miR-143 in the 3′ untranslated region (UTR) of *Elk-1* and for miR-145 in the 3′ UTR of *Myocd* (Supplementary Fig. 6a). Growth signals repress smooth muscle gene expression by displacing Myocd from SRF with Elk-1, a ternary complex factor that acts as a myogenic repressor and an activator of VSMC proliferation[22]. In this system, SRF serves as a platform for myogenic coactivators or corepressors that compete for a common docking site, thereby mediating VSMC phenotypic switching.

To determine whether Elk-1 and Myocd are direct targets of miR-143 or miR-145, respectively, we cloned the 3′ UTR of *Elk-1* or *Myocd*

©2009 Macmillan Publishers Limited. All rights reserved

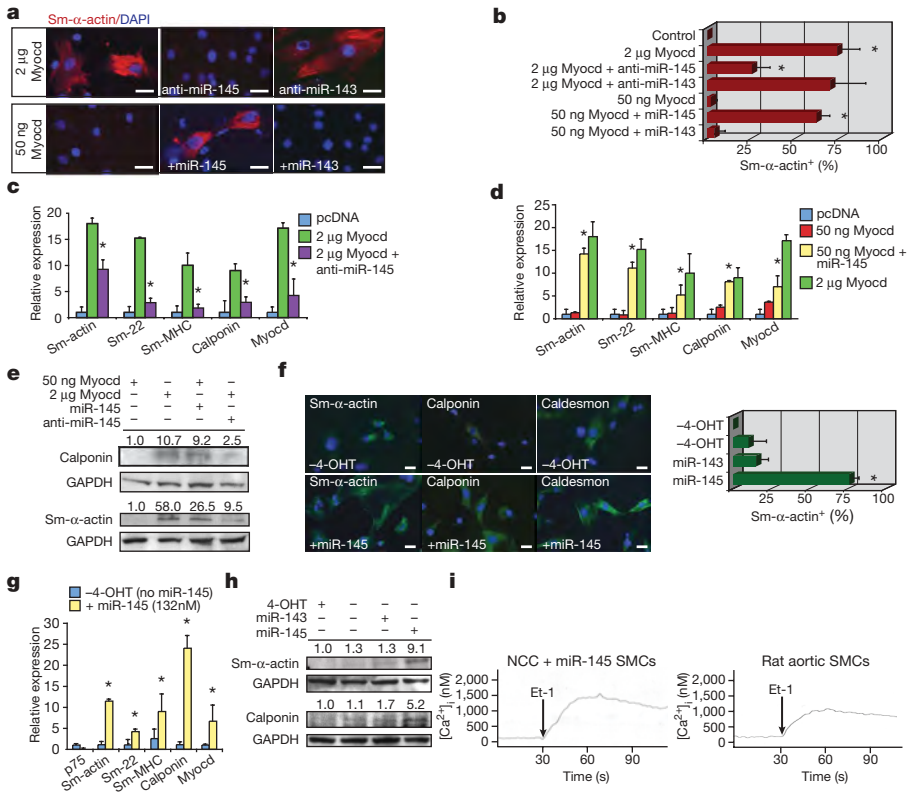


Figure 3 | miR-145 directs vascular smooth muscle cell fate. **a**, Immunocytochemistry of 10T1/2 fibroblasts using smooth muscle (Sm) α-actin antibodies (red) under conditions indicated; nuclear stain, DAPI (blue). **b**, Quantification of Sm-α-actin positive cells ($n = 6$). **c**, qPCR of Sm gene expression in 10T1/2 fibroblasts transfected with Myocd with or without anti-miR-145 or **d**, in 10T1/2 fibroblasts transfected with 50 ng Myocd with or without miR-145 ($n = 5$). **e**, Western blot of calponin and Sm-α-actin. **f**, Left: immunocytochemistry of neural crest stem cells (Joma1.3 NCCs) with or without miR-145 using antibodies indicated (green); tamoxifen (4-OHT) was removed to allow differentiation. Right: quantification of percentage Sm-α-actin$^+$ cells (Joma1.3 NCCs) relative to total DAPI$^+$ nuclei (blue) ($n = 6$). **g**, qPCR of Sm gene expression in Joma1.3 NCCs with miR-145 expression ($n = 5$); p75 is a marker of undifferentiated neural crest cells. **h**, Western blot of Sm-α-actin and calponin. **i**, Calcium flux $[Ca^{2+}]_i$ in VSMCs derived from NCCs or rat aortic VSMCs in response to endothelin-1 (Et-1) stimulation at 30 s. Error bars, s.d.; $^*P < 0.05$.

into the 3′ UTR of a cytomegalovirus (CMV)-driven luciferase reporter. In the presence of the *Elk-1* 3′ UTR, miR-143 repressed luciferase activity; this repression was diminished by mutation of one of the two miR-143 binding sites (Fig. 4a). The addition of an antagomir to inhibit miR-143 in the A10 rat aortic VSMC line resulted in upregulation of Elk-1 protein, but not mRNA, consistent with translational repression of Elk-1 by miR-143 (Fig. 4b, Supplementary Fig. 6b). Furthermore, inhibition of miR-143 caused a doubling of the proliferative rate of VSMCs, demonstrating the function of miR-143 in negatively regulating VSMC proliferation (Fig. 4c).

The presence of putative miR-145 binding sites in the *Myocd* 3′ UTR seemed counter to the observed effects of miR-145 in potentiating Myocd's reprogramming effects. When we cloned the *Myocd* 3′ UTR into a CMV-driven luciferase vector and introduced this into Cos cells, the constitutive luciferase activity decreased greater than 100-fold. Surprisingly, introduction of miR-145, but not miR-143, with the luciferase vector in Cos cells resulted in relief of the repression and an ~250-fold increase in luciferase activity compared to the CMV-luciferase-*Myocd* 3′ UTR-luciferase vector alone (Fig. 4d). The increase in luciferase activity was largely lost on mutation of the miR-145 binding site in the *Myocd* 3′ UTR (Fig. 4d). In contrast, introduction of the same CMV-luciferase-*Myocd* 3′ UTR reporter did not cause a decrease in baseline luciferase activity in 293T epithelial cells. However, even in these cells, miR-145 consistently increased luciferase activity by ~1.5-fold (data not shown). Although antibodies to detect endogenous Myocd levels by western blot are not available, these findings are consistent with the recent observation that miRNAs can act as translational activators or repressors based on the state of the cell cycle[6]. Although the mechanism for this remains unclear, it will be interesting to determine if miR-145 prevents binding of a repressive RNA-binding protein enriched in Cos cells.

Although miR-145 may result in increased Myocd protein, its effects in potentiating Myocd-induced reprogramming of fibroblasts did not require the presence of its binding site in Myocd's 3′ UTR. The potentiating effects of miR-145 could, however, be through effects on translation of endogenous *Myocd* mRNAs induced by the transfected Myocd protein; alternatively, miR-145 may also promote differentiation through targets independent of Myocd. Indeed, our bioinformatics approach identified potential miR-145 binding sites in several other positive regulators of smooth muscle proliferation, including Kruppel-like factor 4 (Klf4) and Calmodulin kinase II-delta (CamkIIδ). Klf4 is a transcription factor involved in pluripotency[30] that is also rapidly induced in post-injury proliferating VSMCs, where it interacts with enhancers in smooth muscle growth genes, inhibits smooth muscle differentiation genes, and represses Myocd expression[31]. The miR-145 binding site in the 3′ UTR of *Klf4* specifically mediated miR-145-dependent repression in luciferase assays (Fig. 4e and Supplementary Fig. 6c). Furthermore, knockdown of miR-145 in rat A10 VSMCs resulted in an increase in Klf4 protein levels, but no change in *Klf4* mRNA levels (Fig. 4f, Supplementary Fig. 6d). Similarly, a putative binding site in *CamkII-δ*, involved in multiple events including neointimal proliferation[32,33], was validated as a miR-145-repressed target by luciferase and western analysis in VSMCs (Fig. 4g, h and Supplementary Fig. 6e). Numerous predicted targets for both miRNAs that were not validated in luciferase assays are shown (Supplementary Fig. 6f, g).

Consistent with miR-145 repression of genes involved in VSMC proliferation, introduction of miR-145 was sufficient to suppress the

©2009 Macmillan Publishers Limited. All rights reserved

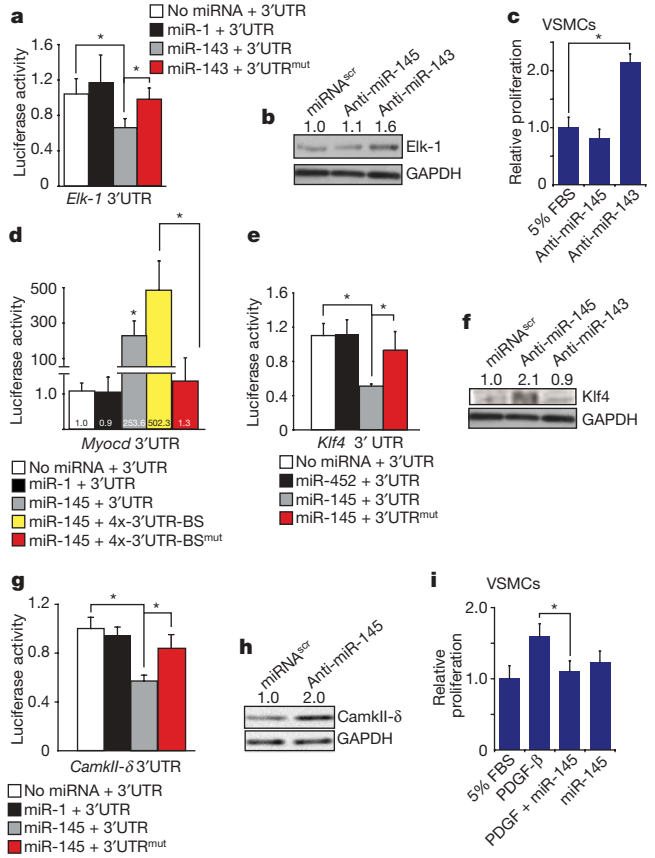


Figure 4 | miR-143 and miR-145 target a network of factors to promote VSMC differentiation and repress proliferation. **a**, Luciferase activity in Cos cells on introduction of *Elk-1* 3′ UTR or mutant 3′ UTR (mut) downstream of a CMV-driven luciferase reporter with indicated miRNAs ($n = 5$). **b**, Elk-1 protein in cell lysates from A10 VSMCs transfected with a scrambled (scr) miRNA or anti-miR-143 or anti-miR-145 assessed by western blot. **c**, Proliferation of VSMCs on inhibition of miR-143 or miR-145 relative to control (5% FBS, $n = 5$). **d**, Luciferase activity in Cos cells with *Myocd* 3′ UTR sequences with indicated miRNAs ($n = 5$). The Myocd binding site (BS) was mutated in the context of a 4× concatemer. **e**, Luciferase activity with WT or mutated *Klf4*-3′ UTR upon introduction of indicated miRNAs ($n = 5$). **f**, Analysis of Klf4 protein in cell lysates from A10 cells transfected with indicated anti-miRs by western blot. **g**, Luciferase activity of WT or mutated *CamkII-δ* 3′ UTR ($n = 5$). **h**, Western analysis for CamkII-δ protein in A10 cells transfected with scr miRNA or anti-miR-145. **i**, Proliferation of VSMCs relative to control ($n = 5$). Error bars, s.d. Densitometry calculation performed by Image J. $^*P < 0.05$.

proliferative response normally induced by platelet-derived growth factor (PDGF-β) in cultured VSMCs (Fig. 4i). In addition, lentiviral-mediated introduction of miR-145 into ligated mouse carotid arteries *in vivo* increased expression of markers of VSMC differentiation (for example, Calponin and Sm-α-actin), as well as Myocd, compared to control-infected injured carotid arteries (Supplementary Fig. 7). These findings suggest that miR-145 promotes VSMC differentiation by directly repressing numerous transcription factors that promote the proliferative state while stabilizing factors that promote the differentiated state of VSMCs (Fig. 5).

Discussion

The ability of miR-145 to efficiently direct VSMC differentiation from multipotent stem cells is the first evidence, to our knowledge, of a miRNA capable of directing the VSMC fate. This is consistent with its early expression in the aorta of developing embryos, as VSMC differentiation is initiated from neural crest and mesodermal progenitors. Once VSMC identity is established in the embryo, the downregulation of miR-143 and miR-145 in the developing embryo

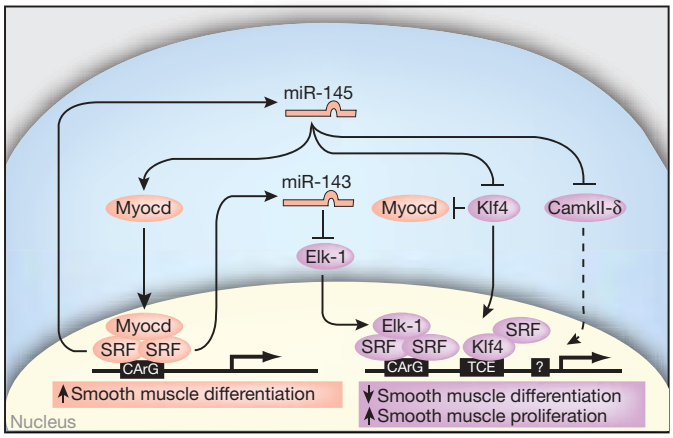


Figure 5 | Model of miR-143 and miR-145 regulation of smooth muscle cell proliferation and differentiation. miR-143 and miR-145 are positively regulated by SRF and function to repress multiple factors that normally promote the less differentiated, more proliferative smooth muscle phenotype (lavender). These include Klf4, which interacts with SRF and also represses Myocd. miR-145 has a positive effect on Myocd activity to concurrently promote the more differentiated smooth muscle phenotype (peach), thereby also functioning to reinforce miR-145 and miR-143 expression. Effects of miR-145 and miR-143 converge on SRF-dependent transcription by regulation of co-activators and co-repressors to dictate the proliferative or differentiated phenotype of VSMCs. Dashed lines indicate indirect effects.

may allow necessary proliferation. Subsequent upregulation postnatally (coincident with the more differentiated VSMC state) and downregulation during neointimal hyperplasia indicate dynamic regulation that may contribute to the oscillating state of VSMCs. The observation that miR-145 is necessary and sufficient for VSMC differentiation raises the possibility that restoration of this miRNA could suppress the smooth muscle hyperplasia observed in vascular injury and atherosclerosis. Furthermore, the multiple targets we identified for miR-143 and miR-145 reveal an elegant mechanism by which this family of miRNAs promotes differentiation and simultaneously represses proliferation of VSMCs by converging on SRF-dependent co-activators and co-repressors (Fig. 5). Our findings suggest miR-145 promotes VSMC differentiation in part by increasing Myocd protein and functioning in a feed-forward reinforcement of its own expression by the SRF-Myocd complex, while miR-143 represses Myocd's competitor, Elk-1. Given the potent effects on differentiation shown here, future studies will determine if restoration of normal levels of miR-143 and miR-145 holds therapeutic value in the setting of vascular disease.

The downregulation of miR-145 in numerous cancers and our findings that it promotes differentiation raises the possibility that miR-145 functions as a pro-differentiation factor in a lineage-specific fashion depending on the cellular context. The targeting of Klf4 supports this notion, as Klf4 is expressed in undifferentiated ES cells and in other less differentiated cell types. Recent evidence indicates that miR-145 represses Klf4 in human ES cells as they begin to differentiate and is required for normal differentiation[34]. As Klf4 is one of four factors that together are sufficient to reprogram human fibroblasts into a pluripotent state (induced pluripotent stem cells)[30], it will be interesting to determine whether inhibition of miR-145 can enhance generation of pluripotent stem cells.

METHODS SUMMARY

Transgenic mice were generated by pronuclear injection and assayed by Bluo-gal (Invitrogen) staining as described[7]. Constructs contained promoter fragments cloned upstream of the pHsp68LacZ reporter. Embryonic hearts were collected at E9.5 from *Islet1-cre* mice[19] crossed with Rosa26-YFP mice[20], and qPCR was performed from YFP^+ cells. Electromobility shift assay (EMSA) was performed as described[35] using oligos corresponding to the conserved SRF-binding sites in

©2009 Macmillan Publishers Limited. All rights reserved

the miR-143/145 enhancer. To identify and validate putative miRNA target genes, we used an in-house automated algorithm[7,10] and cloned the 3′ UTR of each mRNA into the pMiR-Report luciferase reporter. Luciferase activity was measured using the Luciferase Dual-Reporter Kit. To assess miR-143/145 function, rat aortic A10 VSMCs or 10T1/2 fibroblasts were transfected with expression plasmids containing either pre-miR-145 or pre-miR-143, or inhibitors to miR-143 or -145, and then assayed for VSMC-marker gene expression by qPCR, western blot and immunocytochemistry. Myogenic conversion assays were performed as described[36] with 1–2 μg of Myocd. For multipotent stem cell studies, the JoMa1.3 neural crest cell line was maintained as reported[29]. To induce VSMC differentiation, miR-145 was transiently transfected. A10 VSMCs proliferation studies were done as reported[37] using the CellTiter 96TM assay and miRNAs were transfected at varying concentrations; 5 ng ml^{-1} of PDGF-β was added to appropriate wells post-growth arrest. Changes in intracellular calcium concentration ($[Ca^{2+}]_i$) were measured using a calcium fluorescent dye, Indo-1 AM, as described[38] in A10 VSMCs or JoMa1.3 neural crest stem cells transfected with pre-miR-145 and exposed to 10 nM of the calcium agonist, endothelin-1, at 30 s to stimulate calcium flux[38]. For assessment of miR-143 or miR-145 expression during vascular injury, ligated carotid arteries were collected from mice and sectioned, and aortic lesions from apolipoprotein E-null mice fed a western diet were dissected; subsequently, expression was analysed by qPCR and in-situ hybridization.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 February; accepted 10 June 2009.
Published online 5 July 2009.

1. Kloosterman, W. P. & Plasterk, R. H. The diverse functions of microRNAs in animal development and disease. *Dev. Cell* **11,** 441–450 (2006).
2. Calin, G. A. & Croce, C. M. MicroRNA signatures in human cancers. *Nature Rev. Cancer* **6,** 857–866 (2006).
3. Zhao, Y. & Srivastava, D. A developmental view of microRNA function. *Trends Biochem. Sci.* **32,** 189–197 (2007).
4. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–233 (2009).
5. Rajewsky, N. MicroRNA target predictions in animals. *Nature Genet.* **38** (Suppl), S8–S13 (2006).
6. Vasudevan, S., Tong, Y. & Steitz, J. A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318,** 1931–1934 (2007).
7. Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* **436,** 214–220 (2005).
8. Kwon, C., Han, Z., Olson, E. N. & Srivastava, D. MicroRNA1 influences cardiac differentiation in *Drosophila* and regulates Notch signaling. *Proc. Natl Acad. Sci. USA* **102,** 18986–18991 (2005).
9. Chen, J. F. *et al.* The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nature Genet.* **38,** 228–233 (2006).
10. Zhao, Y. *et al.* Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1–2. *Cell* **129,** 303–317 (2007).
11. Ivey, K. N. *et al.* MicroRNA regulation of cell lineages in mouse and human embryonic stem cells. *Cell Stem Cell* **2,** 219–229 (2008).
12. Srivastava, D. Making or breaking the heart: from lineage determination to morphogenesis. *Cell* **126,** 1037–1048 (2006).
13. Kattman, S. J., Huber, T. L. & Keller, G. M. Multipotent flk-1+ cardiovascular progenitor cells give rise to the cardiomyocyte, endothelial, and vascular smooth muscle lineages. *Dev. Cell* **11,** 723–732 (2006).
14. Le Douarin, N. M., Creuzet, S., Couly, G. & Dupin, E. Neural crest cell plasticity and its limits. *Development* **131,** 4637–4650 (2004).
15. Ross, R. The pathogenesis of atherosclerosis: A perspective for the 1990s. *Nature* **362,** 801–809 (1993).
16. Owens, G. K., Kumar, M. S. & Wamhoff, B. R. Molecular regulation of vascular smooth muscle cell differentiation in development and disease. *Physiol. Rev.* **84,** 767–801 (2004).
17. Yoshida, T. & Owens, G. K. Molecular determinants of vascular smooth muscle cell diversity. *Circ. Res.* **96,** 280–291 (2005).
18. Wang, Y., Medvid, R., Melton, C., Jaenisch, R. & Blelloch, R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nature Genet.* **39,** 380–385 (2007).
19. Cai, C. L. *et al.* Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev. Cell* **5,** 877–889 (2003).
20. Srinivas, S. *et al.* Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Dev. Biol.* **1,** 4 (2001).
21. Moretti, A. *et al.* Multipotent embryonic isl1+ progenitor cells lead to cardiac, smooth muscle, and endothelial cell diversification. *Cell* **127,** 1151–1165 (2006).
22. Wang, Z. *et al.* Myocardin and ternary complex factors compete for SRF to control smooth muscle gene expression. *Nature* **428,** 185–189 (2004).
23. Wang, D. *et al.* Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor. *Cell* **105,** 851–862 (2001).
24. Chen, J., Kitchen, C. M., Streb, J. W. & Miano, J. M. Myocardin: a component of a molecular switch for smooth muscle differentiation. *J. Mol. Cell. Cardiol.* **34,** 1345–1356 (2002).
25. Long, X., Bell, R. D., Gerthoffer, W. T., Zlokovic, B. V. & Miano, J. M. Myocardin is sufficient for a smooth muscle-like contractile phenotype. *Arterioscler. Thromb. Vasc. Biol.* **28,** 1505–1510 (2008).
26. Chen, C. Y. & Schwartz, R. J. Recruitment of the tinman homolog Nkx-2.5 by serum response factor activates cardiac alpha-actin gene transcription. *Mol. Cell. Biol.* **16,** 6372–6384 (1996).
27. Ji, R. *et al.* MicroRNA expression signature and antisense-mediated depletion reveal an essential role of MicroRNA in vascular neointimal lesion formation. *Circ. Res.* **100,** 1579–1588 (2007).
28. Krutzfeldt, J. *et al.* Silencing of microRNAs *in vivo* with 'antagomirs'. *Nature* **438,** 685–689 (2005).
29. Maurer, J. *et al.* Establishment and controlled differentiation of neural crest stem cell lines using conditional transgenesis. *Differentiation* **75,** 580–591 (2007).
30. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
31. Liu, Y. *et al.* Kruppel-like factor 4 abrogates myocardin-induced activation of smooth muscle gene expression. *J. Biol. Chem.* **280,** 9719–9727 (2005).
32. House, S. J. & Singer, H. A. CaMKII-delta isoform regulation of neointima formation after vascular injury. *Arterioscler. Thromb. Vasc. Biol.* **28,** 441–447 (2008).
33. Mishra-Gorur, K., Singer, H. A. & Castellot, J. J. Jr. Heparin inhibits phosphorylation and autonomous activity of Ca^{2+}/calmodulin-dependent protein kinase II in vascular smooth muscle cells. *Am. J. Pathol.* **161,** 1893–1901 (2002).
34. Xu, N., Papagiannakopoulos, T., Pan, G., Thomson, J. A. & Kosik, K. S. MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell* **137,** 647–658 (2009).
35. Yamagishi, H. *et al.* Tbx1 is regulated by tissue-specific forkhead proteins through a common Sonic hedgehog-responsive enhancer. *Genes Dev.* **17,** 269–281 (2003).
36. Wang, Z., Wang, D. Z., Pipes, G. C. & Olson, E. N. Myocardin is a master regulator of smooth muscle gene expression. *Proc. Natl Acad. Sci. USA* **100,** 7129–7134 (2003).
37. Yamamoto, M. *et al.* The roles of protein kinase C beta I and beta II in vascular smooth muscle cell proliferation. *Exp. Cell Res.* **240,** 349–358 (1998).
38. Sinha, S. *et al.* Assessment of contractility of purified smooth muscle cells derived from embryonic stem cells. *Stem Cells* **24,** 1678–1688 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Blelloch for *DGCR8*-null EBs; R.J. Schwartz for *SRF*-null ES cells; I. Charo and N. Saederup for RNA from atherosclerotic tissue; J. Maurer for JoMa neural crest cell line; L. Qian and Y. Huang for providing mouse cardiac infarct RNA; C. Tsou for help with calcium flux assays; E. N. Olson for the myocardin expression plasmid; P. Swinton for generation of transgenic mice; J. Fish and C. Miller for histopathology support; S. Ordway and G. Howard for scientific editing; B. Taylor for manuscript preparation. We also thank members of the Srivastava laboratory for discussions. J.M.M. was supported by HL62572 and HL091168 from NHLBI/NIH. D.S. was supported by grants from the NHLBI/NIH and the California Institute for Regenerative Medicine (CIRM) and was an Established Investigator of the American Heart Association. This work was also supported by NIH/NCRR grant C06 RR018928 to the Gladstone Institutes.

Author Contributions K.R.C. and D.S. designed the study and K.R.C. executed or oversaw execution of all experiments; N.T.S. and E.C.B. performed the NCC studies; M.P.W. and K.N.I. performed some expression and stem cell studies and K.N.I. helped supervise the project; A.N.M. provided technical support; T.-H.L. and J.M.M. performed carotid artery ligation studies; S.U.M. isolated YFP^+ progenitor cells and performed some expression studies; J.M.M. assisted K.R.C and D.S. in editing the manuscript; K.R.C. and D.S. wrote the manuscript and D.S. supervised all aspects of the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.S. (dsrivastava@gladstone.ucsf.edu).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Transgenic mice and flow cytometry. Transgenic mice were generated and Bluo-gal staining and histological analyses were performed as described[7]. For promoter analysis, fragments were subcloned into a pHsp68LacZ reporter vector and injected into pronuclei. The 4.2-kb regulatory element corresponds to mouse chr 18:61809195-61813466. *Islet1-cre* mice[19] were crossed with Rosa26-YFP mice[20], embryos were collected at E9.5, and heart and surrounding tissue were dissected, trypsinized, spun at ~300*g* and the pellet was resuspended in PBS and filtered through a 40-μm Millipore membrane. Selection by FACS was based on expression of YFP. YFP^+ cells were collected for RNA preparation.

miRNA qPCR and miRNA *in situ* hybridization. Total RNA was isolated (Trizol, Invitrogen) from mouse E9.5 embryonic hearts and used for quantitative real-time PCR. miRNA *in situ* hybridization analyses were performed as described[39] with the following modifications: paraffin-embedded tissue sections or cryosections were treated for 15 min with Proteinase K, hybridized at 59 °C (miR-145) or 42 °C (miR-143), and final colour development was performed with NBT/BCIP (Roche).

Electromobility shift assay (EMSA). Oligoribonucleotides corresponding to the conserved SRF-binding sites (underlined) and flanking nucleotides in the miR-143/145 enhancer were synthesized (Integrated DNA Technologies) as follows:

SRF binding site: GGGAGCAGCCTTGCCATATAAGGGCAGG; SRF mutant binding site: GGGAGCAGCCTTGCTACCGAAGGGCAGG. EMSA was performed as described[35].

miRNA target prediction. Putative miRNA target genes were identified using an in-house automated algorithm based on empirical miRNA:mRNA interaction data[7,10] that qualifies mRNAs based on (1) complementarity between the seed region of the miRNA and the mRNA 3′ UTR as annotated in RefSeq; (2) identification of an extended binding site; (3) favourable binding affinity between the miRNA and mRNA target site as calculated by RNAhybrid[40]; (4) high free energy in the regions flanking the putative binding site as determined by mFold[41]; (5) absence of stabilizing elements in the binding site; (6) presence of destabilizing elements in the region surrounding and including the possible binding site; and (7) conservation over a number of species.

miR-143 and miR-145 target analyses and expression. A 250-bp fragment encompassing miR-145 was ligated into pSilencer 4.1-CMV (Ambion). A 250-bp fragment containing miR-143 was ligated into pEF-Dest-51 (Invitrogen). The entire 3′ UTR of each mRNA containing predicted miR-143 and/or miR-145 binding sites was cloned into the pMiR-Report luciferase reporter (Applied Biosystems). All assays were performed in quadruplicate in 12-well plates of Cos cells transfected with siPort XP-1 (Ambion). After 24 h, cells were harvested and luciferase activity was measured with the Luciferase Dual-Reporter Kit (Promega). Renilla assays were performed in parallel to normalize for transfection efficiency.

Embryonic stem (ES) cells or embryoid bodies (EBs), A10 cells or differentiated 10T1/2 fibroblasts were harvested in Trizol (Invitrogen) for total RNA isolation. Total RNA (2 μg) from each sample was reversed transcribed with Superscript III (Invitrogen). Taqman primers were used to amplify genes (ABI; primer sequences available upon request). The primers to detect the 1.7-kb miR-143/145 primary transcript were as follows: forward, GCATCTC TGGTCAGTTGGG; reverse, GACCTCAAGAACAGTAT. GAPDH was used as a control. $DGCR8^{null}$ EBs (day 8, D10 EBs) were a gift from R. Blelloch[18]. miRNA qPCR on β-MHC-GFP control EBs, SRF^{null} EBs, or SRF^{null} EBs expressing miR-1 or miR-133 was performed as described above; miR-16 was used as the endogenous control. Each qPCR was performed at least three times; representative results are shown as fold expression relative to undifferentiated ES cells.

Tissue culture. 10T1/2 fibroblasts were maintained at low density (~30% confluence) in DMEM with 10% FBS and were transfected with Lipofectomine 2000 (Invitrogen) and 1–2 μg of full length or smooth muscle isoform of Myocd[36]. Pre-miR-145 sequence containing ~250 bp amplified from genomic DNA was cloned into pSilencer 4.1-CMV vector (Ambion) and pre-miR-143 was cloned into the pEF-Dest-51 vector (Invitrogen). Two days after transfection, media was replaced with differentiation media (DMEM, 2% horse serum). 4–5 days later, further analyses, including immunocytochemistry, western blot and RT–PCR were performed.

A10 VSMCs and Cos cells were maintained in DMEM with 10% FBS. A10 cells were transfected with BlockIT Fluorescent oligo (Invitrogen), miR-143 or miR-145 inhibitor (Dharmacon) or antagomiR (IDT Technologies), or miR-143, 145 mimic (Dharmacon). 24–48 h later, western blot or RT–PCR was performed.

The Joma1.3 neural crest cell line was maintained as reported[29]. NCCs were plated (~7.5×10^5 cells per 10 cm^2) on plastic culture dishes coated with fibronectin, and kept in an undifferentiated state by the addition of 200 nM 4-OHT (Tamoxifen) every 24 h. For differentiation into VSMCs, TGF-β was added 24 h after the last 4-OHT treatment, which was stopped to allow differentiation to take place within 4–6 days. Pre-miR-145 or -miR-143 was transfected in 6-well culture dishes using 10 μl lipofectamine (Invitrogen) at concentrations ranging from 66 nM to 132 nM to induce VSMC differentiation 24 h after removal of 4-OHT.

Proliferation assays. Rat aortic A10 VSMCs proliferation studies were done as reported[37]. Briefly, cells were plated at a density of 5,000 cells per well in 96-well plates containing 5% FBS/DMEM. After plating, miRNAs were transfected at varying concentrations ranging from 20 nM to 240 nM. Twelve hours later, media was washed three times and changed to serum free DMEM with antibiotics. Serum free conditions were maintained for 48 h to allow growth arrest. The medium was then changed to 5% FBS/DMEM and 5 ng ml^{-1} of PDGF-β (R&D Systems) was added to appropriate wells. After 24 h, rates of proliferation were determined using the CellTiter 96TM assay (Promega). Proliferation was measured by the amount of 490 nm absorbance and is directly proportional to the number of living cells. Proliferation was subsequently expressed as absorbance of cells with treatment compared to cells without treatment. Each experiment was done in quintuplicate.

Immunohistochemistry and western blot analysis. Immunostaining was performed using pre-ready mouse anti-smooth muscle actin (Dako, 1A4), 1:500 mouse anti-caldesmon (Abcam, 12B5), and 1:50 rabbit anti-calponin (Chemicon) antibodies and 1:400 Tritc- or Fitc-conjugated goat anti-mouse IgG or goat anti-rabbit IgG (Jackson ImmunoResearch). Myogenic conversion assays were performed as described, and protein lysates collected[36]. Rat aortic A10 cells were collected and assayed using Elk-1, Klf-4, and CamKII-δ antibodies (Cell Signaling).

Calcium flux assays. Calcium studies were performed as described[38]. In brief, rat aortic smooth muscle cells or JoMa1.3 NCCs transfected with pre-miR-145 were grown to 95% confluency, and resuspended in medium containing 1 mg ml^{-1} BSA and 10 mM HEPES to make 1×10^7 cells ml^{-1}. Then the cells were loaded with cell-permeant Indo-1 AM (Invitrogen) for 40 min at 37 °C, and subsequently washed and resuspended with Hank's buffered saline solution containing 1 mg ml^{-1} BSA and 10 mM HEPES. Indo-1 was excited at 350 nm and measured at 410 and 490 nm. The fluorescence intensity ratio (F_{410nm}/F_{490nm}) was calculated using a Hitachi F-2000 fluorescent spectrophotometer and Intracellular Cation Measurement System software, version 1.03 (Hitachi). Cells were exposed to 10 nM of the calcium agonist, endothelin-1 (Sigma) at 30 s to stimulate calcium flux[38].

Mouse vascular injury and atherosclerosis models. Mice that had their left carotid artery ligated were killed 21 days post ligation, fixed and sectioned to obtain cross-sections of the left carotid artery as described[42]; the contralateral right carotid artery was used for control. Pre-packaged pCDH-CMV-MCS-EF1-copGFP control or pre-miR-145 lentivirus (1×10^7 infectious units, System Biosciences) was mixed in 20% pluronic gel and kept cold before application to 12-week-old FVB/NJ mice. Lentiviral-pluronic gel was immediately applied to the external surface of the ligated vessel following arterial injury. Nine days post-ligation, the proximal portion of the injured carotid artery was rapidly removed for total RNA isolation (Trizol) and qPCR (BioRad, MyQ) analysis. 12-week-old apolipoprotein (Apo) E-null mice were fed a Western diet for 4 weeks, and aortic lesions were dissected and collected for RNA analysis.

Statistical analysis. The two-tailed Student's *t*-test, type II, was used for data analysis. $P < 0.05$ was considered significant.

39. Obernosterer, G., Martinez, J. & Alenius, M. Locked nucleic acid-based *in situ* detection of microRNAs in mouse tissue sections. *Nature Protocols* **2**, 1508–1514 (2007).
40. Kruger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* **34**, W451–W454 (2006).
41. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
42. Regan, C. P., Adam, P. J., Madsen, C. S. & Owens, G. K. Molecular mechanisms of decreased smooth muscle differentiation marker expression after vascular injury. *J. Clin. Invest.* **106**, 1139–1147 (2000).

©2009 Macmillan Publishers Limited. All rights reserved

ARTICLES

Architecture and secondary structure of an entire HIV-1 RNA genome

Joseph M. Watts[1], Kristen K. Dang[2], Robert J. Gorelick[5], Christopher W. Leonard[1], Julian W. Bess Jr[5], Ronald Swanstrom[3], Christina L. Burch[4] & Kevin M. Weeks[1]

Single-stranded RNA viruses encompass broad classes of infectious agents and cause the common cold, cancer, AIDS and other serious health threats. Viral replication is regulated at many levels, including the use of conserved genomic RNA structures. Most potential regulatory elements in viral RNA genomes are uncharacterized. Here we report the structure of an entire HIV-1 genome at single nucleotide resolution using SHAPE, a high-throughput RNA analysis technology. The genome encodes protein structure at two levels. In addition to the correspondence between RNA and protein primary sequences, a correlation exists between high levels of RNA structure and sequences that encode inter-domain loops in HIV proteins. This correlation suggests that RNA structure modulates ribosome elongation to promote native protein folding. Some simple genome elements previously shown to be important, including the ribosomal *gag-pol* frameshift stem-loop, are components of larger RNA motifs. We also identify organizational principles for unstructured RNA regions, including splice site acceptors and hypervariable regions. These results emphasize that the HIV-1 genome and, potentially, many coding RNAs are punctuated by previously unrecognized regulatory motifs and that extensive RNA structure constitutes an important component of the genetic code.

Genomes of all single-stranded RNA viruses contain internal structures fundamental to viral replication and host defence evasion. Important viral RNA structures include internal ribosome entry sites, packaging signals, pseudoknots, transfer RNA mimics, ribosomal frameshift motifs, and *cis*-regulatory elements[1,2]. In the human immunodeficiency virus (HIV), RNA structures activate transcription, initiate reverse transcription, facilitate genomic dimerization, direct HIV packaging, manipulate reading frames, regulate RNA nuclear export, signal polyadenylation, and interact with viral and host proteins[2–6]. These RNA regulatory motifs have been identified by focusing on the 5′ and 3′ untranslated regions plus a few internal sequences. Most potential regulatory structures within viral RNA genomes, including in ~85% of the HIV-1 genome, are uncharacterized. This raises the possibility that new categories of RNA structure-mediated regulation remain to be identified.

The HIV-1 genome is primarily a coding RNA and contains nine open reading frames that produce 15 proteins[2,3]. The Gag polyprotein precursor is proteolytically processed to generate the matrix (MA), capsid (CA), nucleocapsid (NC) and p6 proteins. The Gag-Pol polyprotein contains protease (PR), reverse transcriptase (RT) and integrase (IN). The *env* gene encodes a 30-amino-acid signal peptide (SP), gp120 and gp41. Other sequences encode auxiliary proteins (Fig. 1a, grey boxes). Inside virions, HIV genomic RNA is found as a non-covalent dimer, is 5′ capped and 3′ polyadenylated, and is annealed to a host $tRNA^{Lys3}$ molecule[2]. Viral proteins, especially nucleocapsid, chaperone the folding of HIV RNA[7].

Whole-genome structure analysis

To develop an accurate view of RNA structure in the full-length genome, we analysed authentic genomic RNA extracted from HIV-1 virions. Our gentle purification maintained both previously characterized secondary structures and the few known RNA tertiary structures. For example, the host $tRNA^{Lys3}$ was bound to the genome[2] and a pseudoknot in the 5′ untranslated region (UTR)[6,8] remained stably formed. The RNA was sufficiently intact to act as a template for primer extension reactions spanning the entire genome (Supplementary Table 1 and Methods).

High-throughput selective 2′-hydroxyl acylation analysed by primer extension (SHAPE)[6,9–11] was used to chemically interrogate local nucleotide flexibility at 99.4% of the 9,173 nucleotides in the NL4-3 HIV-1 RNA genome. 1-methyl-7-nitroisatoic anhydride (1M7) preferentially acylates conformationally flexible nucleotides at the ribose 2′-OH position[9,10]. The resulting 2′-*O*-adducts are detected as stops to primer extension using fluorescently labelled primers and capillary electrophoresis[6,10] (Fig. 3a) and are quantified by whole-trace Gaussian integration[11] (Fig. 3b). SHAPE measurements are reproducible between independent biological replicates ($R^2 = 0.75$; Supplementary Fig. 1). SHAPE reactivities are highly sensitive to local nucleotide flexibility and disorder, but are insensitive to solvent accessibility[9,12] (Supplementary Fig. 2).

SHAPE reactivities therefore provide direct model-free information about the overall level of structure, or architecture, for any RNA. The median SHAPE reactivity varies markedly across the HIV-1 genome (Fig. 1b, dark blue line). Regions with median reactivities below 0.25 indicate domains with substantial base-paired secondary RNA structure, whereas median SHAPE reactivities of 0.5 and greater indicate regions of largely unstructured nucleotides.

We also assessed HIV-1 genome structure by examining evolutionary information contained in nucleotide and amino acid variation to assign a pairing probability at each nucleotide[13,14]. This algorithm does not use chemical reactivity or thermodynamic information, and thus infers RNA structure using information that is orthogonal to SHAPE.

We identify at least 10 'structured' regions that exhibit both low SHAPE reactivity and high pairing probability (Fig. 1b, compare dark

[1]Department of Chemistry, [2]Department of Biomedical Engineering, [3]Linenberger Cancer Center, [4]Department of Biology, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA. [5]AIDS and Cancer Virus Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702-1201, USA.

©2009 Macmillan Publishers Limited. All rights reserved

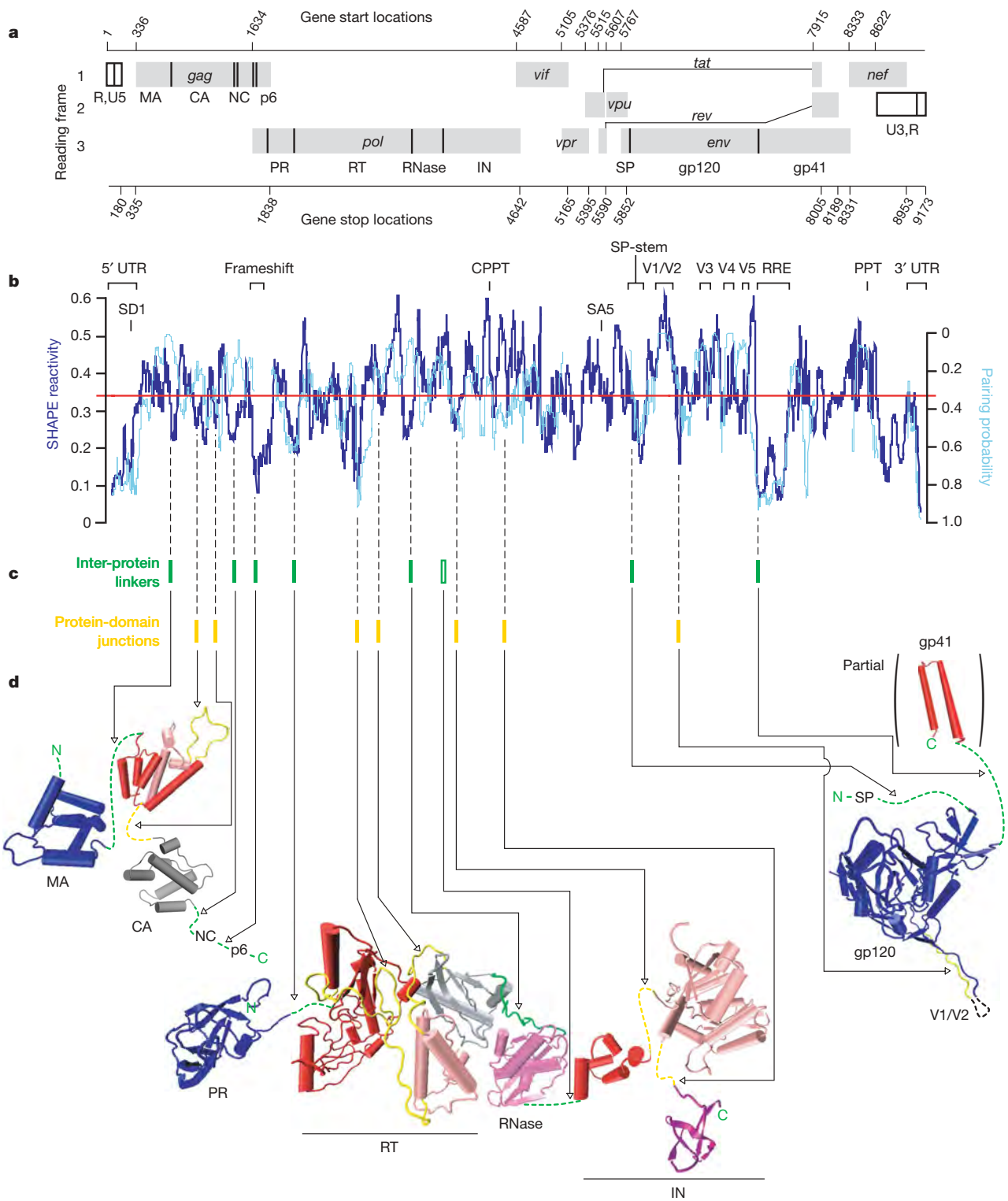


Figure 1 | Organization, extent of RNA structure, and relationship to protein structure for an HIV-1 genome. **a**, HIV-1 genome organization. Protein coding regions are shown as grey boxes; polyprotein-domain junctions are depicted as solid vertical lines. Gene start and end sites are numbered according to NL4-3. CA, capsid; IN, integrase; MA, matrix; NC, nucleocapsid, PR, protease; RT, reverse transcriptase; SP, signal peptide. **b**, Comparison of median SHAPE reactivities (dark blue line) and evolutionary pairing probabilities (cyan line). Medians are calculated using a 75-nucleotide window. The global median (0.34) is depicted as a red line. Pairing probability is not reported for regions encoding overlapping reading frames. PPT, polypurine tract; CPPT, central PPT. **c**, Inter-protein linkers in polyprotein precursors and the unstructured peptide loops that link protein domains are indicated with green and yellow bars, respectively. The single inter-protein linker that is not encoded by a region of highly structured RNA (at the RNase H–integrase junction) is shown with an open green bar. **d**, Comparison of domain structures for the large HIV proteins with the structure of the encoding RNA. Polyprotein linkers are green; inter-domain loops are yellow; folded protein domains are blue, red, light magenta, purple and grey (Supplementary Table 2).

©2009 Macmillan Publishers Limited. All rights reserved

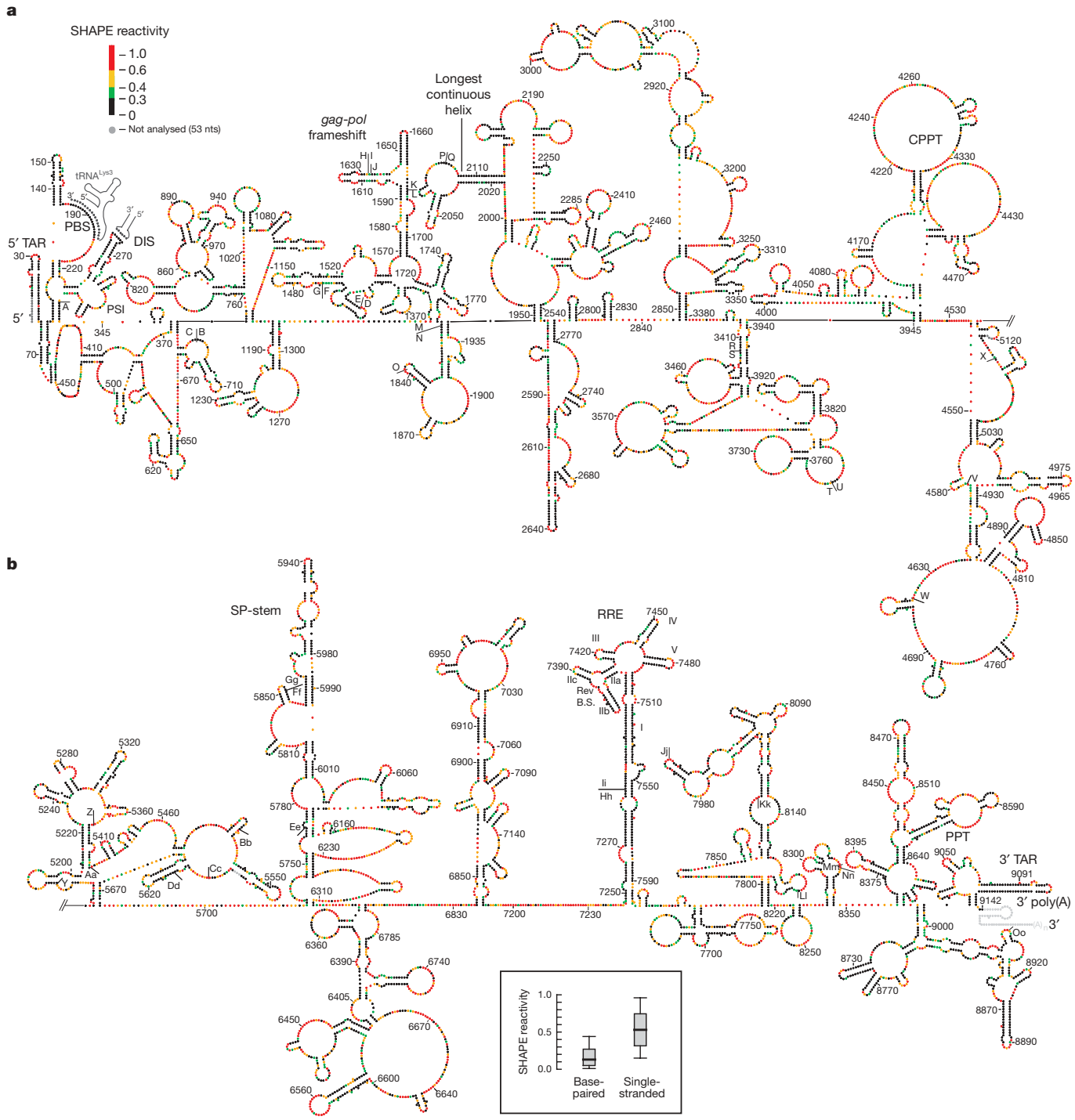


Figure 2 | Structure of the HIV-1 NL4-3 genome. The 5′ (**a**) and 3′ (**b**) genome halves are shown. Nucleotides are coloured by their absolute SHAPE reactivities (see scale in **a**). Every nucleotide is shown explicitly as a sphere; base pairing is indicated by adjacent parallel orientation of the spheres. Protein domains are identified by letters; TF, transframe peptide; nts, nucleotides. Important structural landmarks are labelled explicitly. Full nucleotide identities and pairings are provided in the Supplementary Information (Supplementary Fig. 7). Intermolecular base pairs involving the tRNA^{Lys3} primer and the genomic dimer are shown in grey. Inset shows a box plot illustrating SHAPE reactivities for single-stranded versus paired nucleotides in the model. Median reactivities are indicated by bold horizontal lines; the large box spans the central 50% of the reactivity data.

blue and cyan traces). This group includes the 5′ UTR and the Rev responsive element (RRE), which are known HIV regulatory elements (Fig. 1b). However, most of these highly structured and evolutionarily conserved elements have not been characterized previously. These regions include the protease–reverse transcriptase junction, domains in the reverse transcriptase, integrase, and Vif open reading frames, an element 3′ of the Env signal peptide, and the *nef* 3′-UTR region.

We also identify at least seven ‘unstructured’ regions, extending over 200–600 nucleotides, with high SHAPE reactivities and low

©2009 Macmillan Publishers Limited. All rights reserved

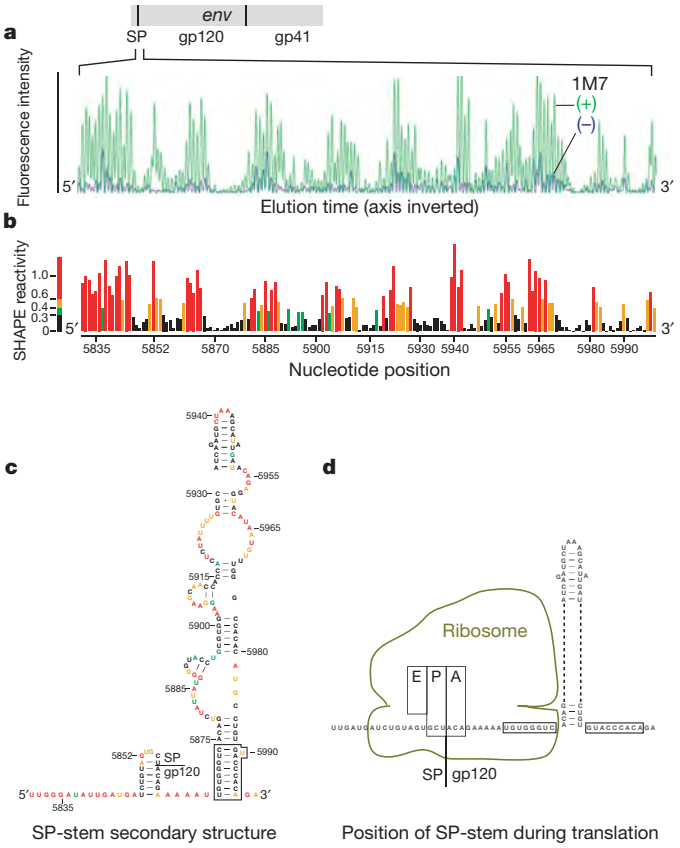


Figure 3 | SHAPE analysis of the signal peptide–gp120 region. a, Processed capillary electrophoresis trace showing intensity versus position for the (+) and (−) reagent reactions. **b**, Histogram of integrated and normalized SHAPE reactivities as a function of nucleotide position. The SHAPE reactivity scale shown here is used consistently throughout this work. **c**, RNA secondary structure model for the signal peptide pause site stem. **d**, Location of the signal-peptide stem relative to the eukaryotic ribosome at the pause site. Base pairs disrupted when the ribosome is at the pause site are boxed.

pairing probabilities. These include the RNase H coding domain, variable domains (V*x*) in gp120, and the polypurine tract (Fig. 1b). On a smaller scale, the consensus sequences for the highly used splice site acceptors are also unstructured (Supplementary Fig. 3). There are four regions of apparent disagreement in the level of RNA structure, having high pairing probabilities and high SHAPE reactivities (one each in the reverse transcriptase, RNase, integrase and gp41 coding regions). This small group may reflect sequence conservation that is not accounted for by the evolutionary model[13], or may form critical structures at an alternative stage of the viral replication cycle.

RNA structure encodes protein structure

We first evaluated whether global RNA genome structure is linked to protein structure. HIV-1 produces three major classes of messenger RNA. The 9 kilobase (kb) class encodes Gag and Gag-Pol and is identical to the packaged genomic RNA analysed here except, as an mRNA, it is not dimerized at its 5′ end[2]. There are very few differences in the SHAPE reactivity of dimeric and monomeric RNAs at the 5′ end of the genome[6]. Thus, genome structures outside of the dimerization region will correlate closely to the mRNA that encodes Gag and Gag-Pol. The most abundant 4 kb *env* mRNA is generated by splicing nucleotide 288 (SD1, the major splice donor) to nucleotide 5522 (termed the SA5 site)[15]. SA5 is followed by an unstructured genome region (Fig. 1a, b). Thus, RNA structures identified in the *env* coding region probably exist in the spliced mRNA that encodes Env. Structures for the 1.8-kb class of mRNAs, which generate Tat and Rev, cannot be predicted using the genomic RNA because discontinuous segments are joined in the final mRNA.

The Gag, Gag-Pol and Env polyprotein precursors are synthesized roughly as beads on a string, and the constituent proteins are liberated by proteolytic cleavage[2,3] (Fig. 1a, d). Eight inter-protein peptides link the HIV proteins (Fig. 1c, green bars). The RNA sequences that encode these spacer peptide linkers in Gag (at the matrix–capsid, capsid–nucleocapsid and nucleocapsid–p6 junctions), Pol (protease–reverse transcriptase and reverse transcriptase–RNase H junctions) and Env (signal peptide–gp120 and gp120–gp41 junctions) all (except the RNase–integrase junction) have SHAPE reactivities that are much lower than the median (Fig. 1b). RNA sequences that encode these inter-protein peptide linkers are more highly structured than 95.2% of randomly selected regions in the genome (Supplementary Fig. 4a).

Domains in the individual HIV-1 proteins—capsid, reverse transcriptase and integrase—are also linked by unstructured peptide elements, and each domain junction is encoded by an RNA region of low SHAPE reactivity (compare yellow bars in Fig. 1c with dark blue trace in Fig. 1b). Protein loops encoded by RNA regions with low SHAPE reactivity include the cyclophilin loop and the linker between the amino- and carboxy-terminal domains in capsid, both loops that link independently folded domains in reverse transcriptase, and the eight and nine amino acid loops linking the three domains in integrase (Fig. 1d, in yellow). These protein-domain junctions are more highly structured than 88.9% of randomly selected equivalent-length regions in the genome (Supplementary Fig. 4b).

In contrast to the other large HIV proteins, domains in gp120 (termed inner, outer and bridging sheet) are not structurally autonomous. The C-terminal 35 residues of gp120 weave from the outer to the inner domain, and the bridging sheet is comprised of residues that are 315 positions distant[16]. Junctions between domains in gp120 are also not encoded by highly structured RNA, suggesting that gp120 folding is not linked to RNA structure in the same way as for other HIV proteins because its constituent domains are not structurally independent.

The recurring pattern of structure, conspicuously located near or after autonomously folding protein coding domains, is consistent with a model in which HIV protein structure is encoded in its RNA at two distinct levels. The first is the linear relationship between RNA and protein primary sequences. In the second level, higher-order RNA structure directly encodes protein tertiary structure, because unstructured protein loops are derived from highly structured RNA elements. Many proteins appear to fold during translation[17], highly structured RNA slows and causes ribosomal pausing during translation[18,19], and changes in the extent of local RNA structure modulate protein activity[20]. Together, these observations suggest that attenuation of ribosome elongation by highly structured RNA at protein-domain junctions facilitates native folding of HIV proteins by allowing time for domains to fold independently during translation.

This model makes the clear prediction that ribosome pause sites should occur preferentially in the highly structured regions of an HIV-1 RNA that encode protein junctions. We tested this idea using a toeprinting experiment, in which ribosome processivity is inhibited by cycloheximide and sites preferentially occupied by the ribosome are detected as stops to primer extension in an *in vitro* translation reaction[21]. Ribosome pause sites are statistically overrepresented at the matrix–capsid and capsid–nucleocapsid junctions in Gag and at the sequences encoding the cyclophilin loop in capsid (Supplementary Fig. 5). Conversely, ribosome pause sites are underrepresented in flanking, but unstructured, regions of the HIV RNA ($P = 0.018$). These experiments thus strongly support the model that mRNA structure over a region spanning 60–100 nucleotides specifically modulates ribosome processivity at protein-domain junctions.

RNA secondary structure model for HIV-1

Comprehensive SHAPE reactivity information can also be used to determine a nucleotide-resolution secondary structure model for the entire NL4-3 HIV-1 genome (Fig. 2). SHAPE reactivities are converted

©2009 Macmillan Publishers Limited. All rights reserved

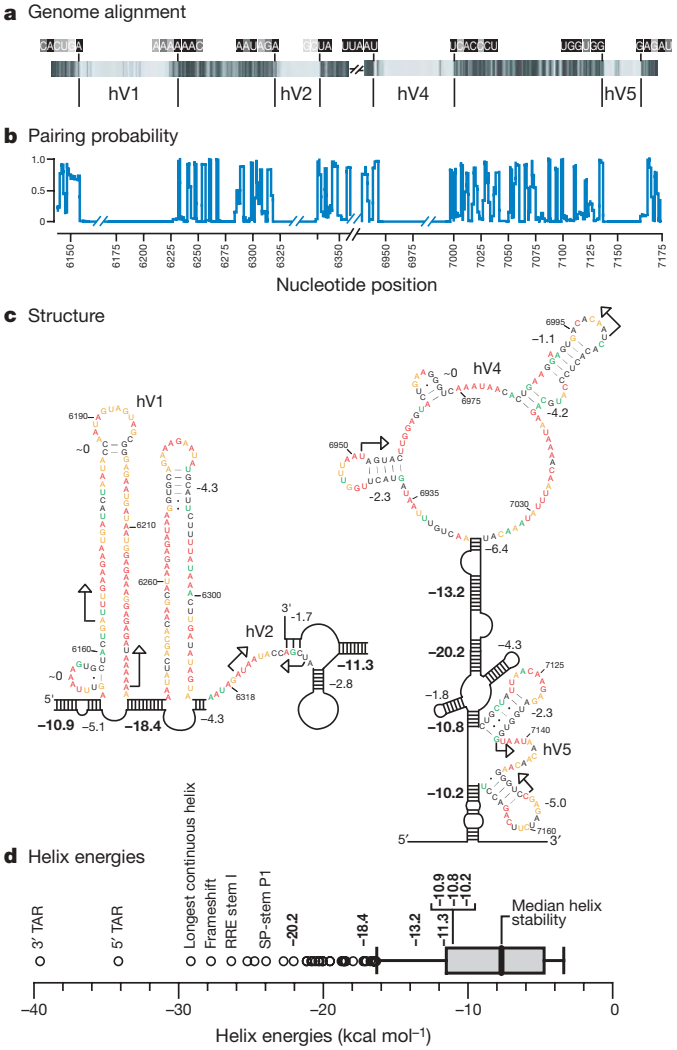


Figure 4 | RNA structure in Env hypervariable regions. **a**, Schematic sequence alignment for group M reference sequences[14] at the Env hypervariable regions (hV1, hV2, hV4 and hV5). Nucleotides are represented as vertical bars; light grey and black indicate low versus universal conservation, respectively. **b**, Evolutionary pairing probabilities. Breaks indicate extensive nucleotide insertions and deletions among the group M consensus sequences. **c**, RNA structures at the hypervariable coding regions hV1, hV2, hV4 and hV5. Calculated free energies are shown for each helix (in kcal mol^{-1}); energies for anchoring helices proposed to function as structural insulators are emphasized in bold. **d**, Distribution of helix stabilities in the HIV genome shown in a box plot representation. Whiskers illustrate 1.5-times the interquartile range, and circles emphasize helices of exceptionally high stability. Free-energy changes for proposed insulating helices are in bold; other significant helices are labelled.

to free-energy change terms and used to constrain a thermodynamic folding algorithm[22,23]. The final result is a thermodynamically favoured structural model highly reflective of the experimental SHAPE data, at single nucleotide resolution. For example, most nucleotides assigned to single-stranded regions are reactive towards SHAPE (Fig. 2, red, orange and green nucleotides), whereas base-paired nucleotides are predominantly unreactive (Fig. 2, black nucleotides and inset). For a full discussion of SHAPE-directed RNA folding and the fundamental correctness of this model, see the Methods.

The HIV-1 genome is less structured than ribosomal RNA but, similarly, contains independent RNA folding domains that extend from the overall genomic backbone. These domains include both small stem-loops plus roughly 21 large and complexly folded structures (Fig. 2). Although many genome regions are highly structured, only seven helices span a complete turn of an 11-base pair (bp) RNA duplex. The largest paired region, devoid of bulges, is the structured RNA element that bridges the coding junction between the reverse transcriptase and RNase H folding domains (Fig. 1). This helix is 19-bp long, contains a non-canonical G-A base pair (Fig. 2a, nucleotides 2015–2033/2103–2121), and is thus shorter than the 30-bp length competent to induce the interferon response[24].

The HIV-1 genome structural model provides a robust starting point for identifying previously unrecognized functional elements and long-range RNA interactions. SHAPE reactivities describe a well-formed stem 3′ to the signal-peptide coding region in the Env protein (Fig. 3c). This stem (the signal-peptide stem) is evolutionarily conserved (Fig. 1b), reinforcing an important biological role. The signal recognition particle (SRP) binds the nascent Env signal peptide and translocates the cytoplasmic ribosome elongation complex to the rough endoplasmic reticulum, where translation of gp120 and gp41 continue[25].

RNA-induced translational pausing occurs as the ribosome unwinds highly structured RNA, typically located 6–7 nucleotides downstream of the A-site[18]. The signal-peptide stem will be exactly in this conformation when the final tRNA^{Ala} from the signal peptide and the first tRNA^{Thr} of gp120 are in the P- and A-sites (Fig. 3d, boxed nucleotides). We infer that ribosomal attenuation or pausing at the signal-peptide stem provides more time for SRP recruitment and subsequent translocation of the elongation complex to the endoplasmic reticulum.

The SHAPE-constrained secondary structure is also informative for previously identified regulatory motifs. In HIV-1, *pro* and *pol* gene products are translated when the ribosome undergoes a −1 register shift from the *gag* to the *pol* reading frames. Frameshifting occurs at a slippery sequence (UUUUUUA) and is enhanced by a downstream RNA structure. These elements are typically drawn as a single-stranded slippery sequence and a 12-bp stem-loop[26]. Direct analysis of intact genomic RNA shows that the *gag-pol* frameshift signal is one component (identified here as P3) of a three-helix structure (Fig. 2 and Supplementary Fig. 6a). The slippery sequence pairs to form one of the three helices (P2). These two helices are stabilized by an anchoring helix (P1) that creates this discrete structural element (Supplementary Fig. 6a). This three-helix junction structure is conserved among HIV-1 group M sequences (Supplementary Fig. 6b).

Most RNA viruses require a complex pseudoknotted structure to induce ribosomal frameshifting[27]. The three-helix junction may function, in part, to slow translation before the ribosome encounters P3, facilitating the prerequisite pause necessary for frameshifting. The three-helix junction model may also explain why changing the slippery site to sequences that allow alternative tRNA pairing and enhance frameshifting in other RNA viruses eliminates frameshifting in HIV-1 (ref. 28). In the SHAPE-directed model, changes to the slippery sequence compromise base pairing in the conserved P2 helix (Supplementary Fig. 6).

Unstructured motifs and insulator helices

Analysis of the HIV-1 genome structure supports a role for RNA structures in sequestering unstructured regions. Five variable domains (V1–V5; see Fig. 1a, b) in the Env surface protein, gp120, account for much of the genetic diversity in HIV-1 (ref. 14) and are a critical component of the viral host evasion strategy. Four of these domains are hypervariable (hV1, hV2, hV4 and hV5) and exhibit large amino acid insertions and deletions between viral isolates[14].

Sequences encoding hypervariable domains are internally unstructured and are bordered by evolutionarily conserved and stable RNA structures (Fig. 4a, b). For example, hypervariable region hV1 is encoded by RNA sequences with high SHAPE reactivities and is flanked by two stable helices (with free energies of −10.9 and −18.4 kcal mol^{-1}, Fig. 4c). Similar patterns are evident in the other hypervariable regions (Fig. 4c). Some hypervariable regions, especially hV4, contain internal helices with non-trivial free energies; however, these helices are not evolutionarily conserved (Fig. 4b)

©2009 Macmillan Publishers Limited. All rights reserved

and are much less stable than the flanking helices that have stabilities in the 10–20 kcal mol^{-1} range (Fig. 4c). These helices are also highly stable relative to the distribution of duplex stabilities over the entire genome (Fig. 4d).

Collectively, these data suggest that RNA sequences encoding length polymorphisms in *env* are segregated from the rest of the genome by stable helices that function as structural insulators. The observed organization of hypervariable regions is thus well suited, first, to accommodate extensive substitutions, insertions or deletions, and second, to prevent these regions from forming non-functional base-pairing interactions with adjacent regulatory motifs, which include the 3′ splice site acceptors and the RRE.

Perspective

Structural analysis of a complete HIV-1 genome reveals that this RNA is punctuated by previously unrecognized, but readily identifiable and evolutionarily conserved, RNA structures. Most genome regions with low SHAPE reactivities are associated with a regulatory function (Fig. 1). SHAPE may be generally useful for identifying new regulatory elements in large RNAs. All of these elements represent hypotheses and starting points that we hope will stimulate further detailed examination. Our discovery that the peptide loops that link independently folded protein domains are encoded by highly structured RNA indicates that these and probably other mRNAs encode protein structure at a second level beyond specifying the amino acid sequence. In this view, higher-order RNA structure directly encodes protein structure, especially at domain junctions. The extraordinary density of information encoded in the structure of large RNA molecules (Figs 1, 2 and 4d) represents another level of the genetic code, one which we understand the least at present. This work makes clear that there is much to be discovered by broad structural analyses of RNA genomes and intact mRNAs.

METHODS SUMMARY

Full length HIV-1 genomic RNA was gently purified from NL4-3 virions (GenBank accession AF324493). The RNA was equilibrated in a native buffer (50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM $MgCl_2$) at 37 °C for 15 min and treated with 1M7 (ref. 10). Sites of 2′-hydroxyl modification were identified over read lengths spanning several hundred nucleotides using 31 primer extension reactions resolved by fluorescence-detected capillary electrophoresis[6,11]. Pairing probabilities were determined using RNA-Decoder[13] and secondary structure models were developed by incorporating SHAPE reactivities as a pseudo-free-energy change term, in conjunction with nearest-neighbour parameters, in an accurate thermodynamics-based prediction algorithm[22,23].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 May; accepted 22 June 2009.

1. Cann, A. J. *Principles of Molecular Virology* Ch. 2–5 (Elsevier, 2005).
2. Coffin, J. M., Hughes, S. H. & Varmus, H. E. *Retroviruses* (Cold Spring Harbor Laboratory Press, 1997).
3. Frankel, A. D. & Young, J. A. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67,** 1–25 (1998).
4. Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5′ region of the HIV-1 genome. *J. Mol. Biol.* **336,** 369–379 (2004).
5. Goff, S. P. Host factors exploited by retroviruses. *Nature Rev. Microbiol.* **5,** 253–263 (2007).
6. Wilkinson, K. A. *et al.* High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.* **6,** e96 (2008).
7. Levin, J. G., Guo, J., Rouzina, I. & Musier-Forsyth, K. Nucleic acid chaperone activity of HIV-1 nucleocapsid protein: critical role in reverse transcription and molecular mechanism. *Prog. Nucleic Acid Res. Mol. Biol.* **80,** 217–286 (2005).
8. Paillart, J. C., Skripkin, E., Ehresmann, B., Ehresmann, C. & Marquet, R. *In vitro* evidence for a long range pseudoknot in the 5′-untranslated and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem.* **277,** 5995–6004 (2002).
9. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127,** 4223–4231 (2005).
10. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129,** 4144–4145 (2007).
11. Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M. & Giddings, M. C. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* **14,** 1979–1990 (2008).
12. Gherghe, C. M., Shajani, Z., Wilkinson, K. A., Varani, G. & Weeks, K. M. Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S^2) in RNA. *J. Am. Chem. Soc.* **130,** 12244–12245 (2008).
13. Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P. & Hein, J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **32,** 4925–4936 (2004).
14. Leitner, T. *et al. HIV Sequence Compendium* (Theoretical Biology and Biophysics Group, 2005).
15. Purcell, D. F. & Martin, M. A. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J. Virol.* **67,** 6365–6378 (1993).
16. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393,** 648–659 (1998).
17. Komar, A. A. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* **34,** 16–24 (2009).
18. Farabaugh, P. J. Programmed translational frameshifting. *Microbiol. Rev.* **60,** 103–134 (1996).
19. Wen, J. D. *et al.* Following translation by single ribosomes one codon at a time. *Nature* **452,** 598–603 (2008).
20. Nackley, A. G. *et al.* Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* **314,** 1930–1933 (2006).
21. Hartz, D., McPheeters, D. S., Traut, R. & Gold, L. Extension inhibition analysis of translation initiation complexes. *Methods Enzymol.* **164,** 419–425 (1988).
22. Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA* **101,** 7287–7292 (2004).
23. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure prediction. *Proc. Natl Acad. Sci. USA* **106,** 97–102 (2009).
24. Kim, D. H. *et al.* Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nature Biotechnol.* **23,** 222–226 (2005).
25. Stein, B. S. & Engleman, E. G. Intracellular processing of the gp160 HIV-1 envelope precursor. Endoproteolytic cleavage occurs in a cis or medial compartment of the Golgi complex. *J. Biol. Chem.* **265,** 2640–2649 (1990).
26. Wilson, W. *et al.* HIV expression strategies: ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. *Cell* **55,** 1159–1169 (1988).
27. Giedroc, D. P., Theimer, C. A. & Nixon, P. L. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298,** 167–185 (2000).
28. Biswas, P., Jiang, X., Pacchia, A. L., Dougherty, J. P. & Peltz, S. W. The human immunodeficiency virus type 1 ribosomal frameshifting site is an invariant sequence determinant and an important target for antiviral therapy. *J. Virol.* **78,** 2082–2087 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project was supported by the US National Institutes of Health (AI068462 to K.M.W.) and by the National Cancer Institute, under contracts N01-CO-12400 and HHSN261200800001E (to R.J.G. and J.W.B.). J.M.W. was supported as a Fellow of the UNC Lineberger Cancer Center and a National Institutes of Health (NIH) Kirschstein Postdoctoral Fellowship. R.S. and K.K.D. were supported by NIH grants AI44667 and T32 AI07419, respectively. We are indebted to D. Mathews and J. Low for assistance with the RNA structure program and genome secondary structure analysis, respectively. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations indicate endorsement by the US Government.

Author Contributions J.M.W., R.J.G. and K.M.W. conceived of and designed the HIV-1 genome structure analysis project. J.M.W. and K.M.W. analysed and interpreted the HIV SHAPE structure information. K.K.D., R.S. and C.L.B. designed and performed the bioinformatic pairing probability analysis. J.M.W., R.J.G. and C.W.L. performed the experiments. J.M.W., C.L.B. and K.M.W. performed the statistical analyses. J.W.B. produced and purified HIV-1 virions. J.M.W. and K.M.W. wrote the manuscript with contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.M.W. (weeks@unc.edu).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Virus production. HIV-1 strain NL4-3 (group M, subtype B) was used to infect a non-Hodgkin's T cell lymphoma cell line (a modified version of the SupT1 cell line, which was a gift from J. Hoxie)[29]. The virus-containing inoculum for infecting SupT1 cells was generated by $CaPO_4$/DNA coprecipitation[30] and subsequent transfection of pNL43 (NIH AIDS Research and Reference Reagent Program; GenBank accession AF324493) into 293T cells[31]. HIV-1 virions were purified as described[32] except cells were removed using a Millipore Opticap XL-5.0 micron filter. The total protein and CAp24 yields were 20.7 mg and 2.3 mg, on the basis of total protein (BioRad DC protein assay) and HPLC with subsequent amino acid analysis assays, respectively.

Virions were purified from cellular debris by subtilisin treatment and centrifugation through a sucrose cushion. Concentrated virions (in 19 ml, corresponding to 19 l of infected cell-culture supernatant) were digested with subtilisin (1 mg ml^{-1}, in 20 mM Tris (pH 8.0), 1 mM $CaC1_2$, 37 °C, 18 h; stopped by the addition of 5 μg ml^{-1} phenylmethylsulphonyl fluoride[33]). The resulting solution contained digested cellular proteins and viral particles free of surface proteins. The sample was centrifuged through a cushion of 20% (w/v) sucrose in PBS (Beckman SW41 rotor, 235,000*g*, 2 h, 4 °C); supernatant was carefully removed, and residual sucrose in the pellet was removed by overlaying PBS and repeating the centrifugation step (1 h at 4 °C).

RNA extraction. The key features of this protocol are that genomic RNA was gently extracted from purified virions in the presence of buffers containing monovalent and divalent ions, consistent with maintaining RNA secondary and tertiary structure. The HIV genomic RNA was not denatured by heat, chemical denaturants, magnesium chelation, or removal of monovalent cations during this process. Subtilisin-treated virions were suspended in virion lysis buffer (VLB; 50 mM HEPES (pH 8.0), 200 mM NaCl, and 3 mM $MgCl_2$) and lysed with 1% (w/v) SDS and 100 μg ml^{-1} proteinase K (~22 °C, 30 min). The digest was extracted three times with phenol/chloroform/isoamyl alcohol (25:24:1, pre-equilibrated with VLB), followed by two extractions with pure chloroform. Quantitative reverse-transcriptase PCR was used to quantify viral RNA yields against a standard curve[34–36]. The total yield from 19 l of infected cells was 97.2 pmol HIV-1 genomic RNA. The aqueous layer (3.6 ml) was brought to 300 mM NaCl and precipitated with 70% (v/v) ethanol. Retroviral genomes commonly contain single-stranded breaks[2]. Approximately 30% of our genomic RNA was intact, as judged by visualization in agarose/formaldehyde gels; nicks in the remaining 70% seemed to be roughly randomly distributed on the basis of direct visualization of the genomic RNA and from the continuity of our primer extension reactions (see Supplementary Table 1).

RNA modification. The RNA pellet containing 97.2 pmols of HIV-1 genomic RNA was dissolved in 880 μl of modification buffer (50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM $MgCl_2$) and incubated at 37 °C for 15 min. Then, 405 μl of the solution was added to 45 μl pre-warmed (37 °C) 1M7 (in dimethylsulphoxide (DMSO))[10] or to DMSO. After 4 min, 45 μl of 50 mM EDTA (pH 8.0) were added to each tube. The reactions were divided into 11-μl aliquots and precipitated with ethanol.

Primer synthesis. Primers were designed with the aid of OligoWalk, part of the RNAstructure software package[22] (available for download at http://rna.urmc.rochester.edu/) (Supplementary Table 1). Primers were required to be 20–22 nucleotides in length, have high melting temperatures and low self-annealing energies, and preferably end with a 3′ G or C. Only 2 out of 31 primers required redesign, giving OligoWalk a 94% success rate. Primers were synthesized to contain a 5′ six carbon linker terminating in a primary amine (IDT). The amine-tethered DNA primers (1 μl; 25 μg ml^{-1}) were labelled with one of four fluorophores (5-FAM, 6-JOE, 6-TAMARA or 5-ROX; AnaSpec) using *N*-hydroxysuccinimide chemistry (3 μl NHS-coupled dye (20 mg ml^{-1} in DMSO) in 0.1 M $NaBO_3$-HCl (pH 8.5); ~22 °C, 3 h). Labelled primers were precipitated with ethanol, purified on a denaturing gel (20% 29:1 acrylamide/bis-acrylamide, 7 M urea, 1× TBE), recovered by passive elution in water, precipitated (300 mM NaCl, 2.5 vol ethanol, 1 vol isopropanol), pelleted, and dissolved in water. Spectrophotometric measurements indicated labelling was ~90–95% efficient as determined by the [dye]/[DNA] ratio.

Primer extension. RNA pellets (1 pmol) were dissolved in 10 μl 0.5× TE (5 mM Tris (pH 8.0), 0.5 mM EDTA) and mixed with 3.0 μl of 0.4 μM primer. The (+) and (−) 1M7 reagent reactions were labelled with JOE and FAM, respectively. Primers were annealed to the RNA by heating to 65 °C for 5 min and 45 °C for 2 min, and then placed on ice. Six microlitres of reverse transcriptase mix[37] (SuperScript III, 5× buffer, DTT, dNTPs; Invitrogen) was added to each tube and incubated for 10 s at 45 °C, 5 min at 52 °C, 5 min at 65 °C, and cooled to 4 °C. Sodium acetate (pH 5.2; 2.0 μl at 3 M) was added to each tube, (+) and (−) 1M7 tubes were combined, and 120 μl of ethanol was added to precipitate the cDNA products. The reactions were pelleted, washed with 70% ethanol, and dissolved in 10 μl deionized formamide.

Sequencing. Dideoxy sequencing reactions (GenomeLab Methods Development Kit; Beckman) were performed using plasmids pDR0 and pDR25 (containing partial NL4-3 sequences), and primers were labelled with TAMARA and ROX. Primer sequences were identical to those in Supplementary Table 1 except primer 31, the sequence (5′-CTGCAACCTCTACCTCCTG GGTGCTAGAG-3′) of which annealed to the plasmid rather than the poly(A) RNA sequence in the genomic RNA.

Capillary electrophoresis. cDNA fragments were resolved by capillary electrophoresis[6,10] (Applied Biosystems AB3130 instrument). Samples were injected at 1.2 kV for 16 s into a 36-cm capillary containing POP7 (ABI) and subjected to electrophoresis for 25 min at 15 kV. The fluorescence detector was initially calibrated with 5-FAM, 6-JOE, 6-TAMARA and 5-ROX using fluorescent markers with fragment lengths of 242 (5-FAM), 206 (6-JOE), 188 (6-TAMARA) and 155 (5-ROX) nucleotides. Fragments were generated by linear amplification of HindIII-digested plasmid pUC18 using primers with the sequences 5′-CAGAGCAGATTGTACTGAGAG-3′, 5′-GTGAAATACCGCAC AGATGC-3′, 5′-GCGTAAGGAGAAAATACCGCATC-3′ and 5′-CGCCATTC AGGCTGCGCAACTG-3′, labelled with 5-FAM, 6-JOE, 6-TAMARA and 5-ROX, respectively. Fluorescent spectral overlap based on this DNA ladder was calibrated using AB3130 software.

Data processing. Raw electropherograms, containing fluorescence intensity versus elution time information, were converted to normalized SHAPE reactivities using ShapeFinder[6,11,23] (available for download at http://bioinfo.unc.edu). The ShapeFinder software aligns the (+) and (−) reagent traces to the two dideoxy nucleotide sequencing ladders, corrects for signal decay[38], and performs a whole-channel Gaussian integration[11] to quantify all individual peak areas (see Fig. 3a). Only 11 of the 9,173 nucleotides in the NL4-3 RNA genome had high background and were therefore excluded from analysis. Data sets were normalized to a scale such that 1.0 represents the average intensity of highly reactive nucleotide positions[6,23]. On this scale, ~95% of integrated intensities for the HIV-1 genome fall between 0 and 1 (see histogram in Fig. 3b). Each primer extension reaction was processed individually. The resulting intensities in regions with overlapping data from different primers correlated closely: reactivity differences were typically less than 0.1 SHAPE unit. Regions with overlapping data accounted for ~25% of the total nucleotide positions and were averaged to generate the final data set spanning the entire NL4-3 genome.

Toeprinting ribosome pause sites at the matrix–capsid and capsid–nucleocapsid junctions. A double-stranded DNA template to direct synthesis of an mRNA spanning NL4-3 Gag nucleotides 1 to 1795 was generated by PCR. This region encompasses the entire 5′ UTR and most of the *gag* coding region and ends after the three-stem frameshift element. The 5′ primer included a T7 promoter sequence (5′-TAATACGACTCACTAATGGTCTCTCTGGTTAGACCA-3′), and the 3′ primer (5′-GCTAAAGGTTACAGTTCCTTGTC-3′) encoded a stop codon at position 1787. The RNA transcript was capped and polyadenylated (mSCRIPT, Epicentre) and *in vitro* translation was carried out in rabbit reticulocyte extract (Ambion) using ~60 μg of the capped, polyadenylated transcript, 1 μl 1.25 mM L-methionine, 1 μl ^{35}S-methionine (PerkinElmer), 17 μl reticulocyte extract, and 1.25 μl 20× 'medium-salt' translation buffer (Ambion) in a total volume of 26 μl at 37 °C. Cycloheximide was added at 0, 7, 15 or 45 min to arrest translation[21]. Translation reaction aliquots were separated on an 8–16% SDS–PAGE gel (Invitrogen) to confirm production of a protein of the correct length. Sites of ribosome pausing were detected by adding the following to 25 μl of the *in vitro* translation mixture: 1.35 μl 10 mM each dNTP, 2 μl 4.0 μM fluorescently labelled primer (primer 4 or 6 for interrogating the matrix–capsid and capsid–nucleocapsid regions, respectively), 1 μl 200 mM $MgCl_2$, and 2 μl Superscript III (Invitrogen). The translation reaction that was pre-quenched with cycloheximide was taken as background and was resolved using a JOE-labelled primer. The 7, 15 and 45 min time points were resolved using FAM-labelled primers. Primer extension reactions were incubated at 37 °C for 30 min and stopped by the addition of 1 μl 0.5 M EDTA and 400 μl water. The reaction was extracted with phenol:chloroform:isoamyl alcohol (25:24:1, 2×) and chloroform (1×). Four microlitres of this solution, 1 μl of a cDNA sequencing ladder, and 15 μl of formamide were combined, heated to 105 °C for 5 min, and resolved by capillary electrophoresis. Toeprinting traces were processed with ShapeFinder[11] and normalized to a scale in which 1.0 is equal to the mean intensity of the most reactive positions, identically as described above for SHAPE experiments.

RNA secondary structure model. The entire NL4-3 sequence—9,173 nucleotides plus 20 3′ adenosines (representing the poly(A) tail)—was folded using the thermodynamics-based algorithm in RNAstructure[22,23]. SHAPE information was used to constrain secondary structure calculations by incorporating SHAPE reactivities as pseudo free-energy change terms[6,23] using slope and intercept values of 30 and −6, respectively. The maximum distance allowed between any two paired positions was 600 nucleotides. The slope and intercept values are derived from previous parameterization on long RNAs, and the 600-nucleotide

©2009 Macmillan Publishers Limited. All rights reserved

cutoff reflects that 99% of all base pairs in ribosomal RNA occur between nucleotides less than this distance apart[23]. The genome was initially folded as a single (9,193 nucleotides) unit; folding was then fine-tuned by dividing the RNA into five independent folding regions, separated by long stretches of reactive nucleotides that were calculated to be unpaired when the entire genome was folded with SHAPE constraints (NL4-3 residues 1–2844, 2836–5722, 5676–6832, 6807–7791 and 7779–9193). Dividing the genome in this way facilitated model building and prevented the formation of a few poorly supported long-distance pairings between domains. Highly reactive nucleotides at the termini of each region were prohibited from forming base pairs in these region-specific calculations. Helices consisting of a single base pair were removed from the final model and unreactive nucleotides in the primer binding site (183–199) were taken to reflect hybridization with the tRNA primer. The current version of our algorithm does not allow pseudoknots and therefore our HIV-1 structure model (Fig. 2) includes only one, heuristically predicted[6,8], pseudoknot.

Quality of SHAPE-directed model of the entire HIV-1 genome. The algorithm by which SHAPE information is used to create an RNA secondary structure model does not make any specific assumptions about the magnitude of SHAPE reactivity that corresponds to single-stranded versus based-paired regions. Instead, SHAPE reactivities are converted to free-energy change terms and used to constrain a thermodynamic folding algorithm[22,23]. SHAPE information is essential for generating this secondary structure model. Folding the genome by free-energy minimization alone, using a best-of-class algorithm[22,39], results in a structure that is very different from the experimentally supported model. Only 47% of the base pairs in the SHAPE-directed model also occur in the lowest free-energy thermodynamics-only model. The unconstrained thermodynamics-only model is readily shown to be incorrect because many regions with high SHAPE reactivities are assigned as paired in the unconstrained model, and many regions with low SHAPE reactivities are assigned as single-stranded.

Several lines of evidence support fundamental correctness of our working SHAPE-directed HIV-1 genome structural model (Fig. 2). First, SHAPE-directed folding is well validated and predicts the known structures of large RNAs, including 16S ribosomal RNA, with high accuracies (>90%)[10,23]. Second, most nucleotides assigned to single-stranded regions are reactive by SHAPE (Fig. 2, red, orange, and green nucleotides). Conversely, base-paired nucleotides are generally unreactive (Fig. 2, black nucleotides and inset). Thus, the structural modelling faithfully incorporates the experimental data. Third, many single nucleotide bulges are predicted as single reactive positions imbedded in helices with flanking nucleotides that are unreactive towards SHAPE, which speaks to the accuracy at the single nucleotide resolution level (for select examples see Fig. 2, positions 1725, 3376, 4891, 5990, 7431, 7568 and 9091). Fourth, previously characterized HIV RNA structures including the 5′ TAR element, the DIS component of the packaging signal, and the five-stem RRE, serve as positive controls and form structures consistent with previous work[4,40] (Fig. 2). In the case of the *gag-pol* frameshift structure, we note that SHAPE data do not support common alternative proposals for this specific structure, including either a longer bulged stem or a pseudoknot.

Most structures in our current HIV-1 genome model, especially in regions with several closely spaced helices, are extremely well determined, as evidenced by the strong correlation between SHAPE values and base pairing. This correlation is also consistent with benchmarking studies showing that SHAPE reactivities strongly discriminate between base paired and single-stranded nucleotides (Supplementary Fig. 2)[41] and are proportional to the extent of local nucleotide disorder[12]. In contrast, some of the larger loop regions in our model may reflect regions that interconvert between multiple structures[38,42]. Elements that may require future refinement include the precise termini of helices at some multihelix junctions and along the central backbone of the genome structure and the identification of further pseudoknot and long-range interactions.

Calculation of evolutionary base pairing probabilities. RNA-Decoder[13] was used to identify regions in the HIV-1 genome in which the ability to form base pairs is evolutionarily conserved. The program takes a set of grammar parameters, a multiple-sequence alignment, and a phylogenetic tree as input. The output is a pairing probability for each position in the genome, given the phylogenetic tree, alignment, and the grammar structural model. The pairing probability for position i in alignment D is the sum over all stem structural labels (k) of $P(\pi_i = k|M)P(D|\pi,T,M)$, where π is the structure, M is the grammar model parameters, and $P(\pi_i = k|M)$is the posterior probability that position i has the specific structural label k, given the grammar[43], and is calculated by the inside-outside algorithm[44]. In Bayesian terms, $P(\pi_i = k|M)$ is the prior probability of structure π and $P(D|\pi,T,M)$ is the alignment probability, calculated using the Felsenstein algorithm[45]. Pairing predictions were made using an alignment of non-recombinant group M subtype reference sequences obtained from the Los Alamos HIV database[14], with minor manual editing (and excluding subtype G, which is now considered a circulating recombinant form[46]). Codon positions in genome regions encoding more than one protein in overlapping reading frames were defined according to the first open reading frame in the following pairs: *gag-pro*, *pol-vif*, *vpr-vif*, *vpr-tat*, *rev-tat*, *env-vpu*, *env-tat2* and *env-rev2*. Owing to differences in nucleotide content and evolution rate within different genes in the HIV genome, the genome was scanned in two sections, upstream and downstream, that overlapped in the *vif* gene. This allowed use of separate phylogenetic trees for each scan, with branch lengths calculated according to the rates of evolution in each genome region. The phylogenetic tree for the 5′ half was built using the third codon position for the *gag*, *pol* and *vif* genes, and the 5′ non-coding region; the tree for the 3′ half was built on the third positions of *vif*, *vpr*, *rev*, *vpu*, *env* and *nef* genes, and the 3′ non-coding region.

Pairing probabilities were assessed across the entire genome. To accommodate as many pairing interactions as possible, we used a large window size (1,300 nucleotides), and spaced the scans at 300-nucleotide intervals. Pairing probabilities for each scan were combined using the statistical program R[47] taking the maximum pairing probability in overlapping windows. It is important to note that high pairing probabilities identify regions experiencing evolutionary pressure to retain a specific, defined, secondary structure. A low pairing probability, although suggestive of a lack of structure, can also reflect (1) that an additional evolutionary constraint exists that is not accounted for by the evolutionary model, or (2) that natural selection favours folding in general, but not a precise pattern of folding.

Bootstrap analysis of SHAPE reactivities in inter-protein linkers and protein-domain junction. A bootstrap procedure was used to compare the SHAPE reactivities of particular collections of genome elements to the expectation for random genome regions of the same size. For a comparison to a collection of n genome elements, we generated 100,000 bootstrapped samples by randomly choosing n locations from the relevant portion of the genome, and randomly assigning the lengths of the actual genome elements to these n locations. For comparison to the protein-domain junctions, locations were drawn randomly from the entire coding portion of the genome (bases 336–8621). We specified a length of 60 nucleotides for each region. For comparison to the intra-domain loops, locations were drawn randomly from within the domains where loops occur and assigned lengths that reflected loop sizes in the same domain (for example, for the capsid domain, one element of 45 base pairs was drawn from within bases 732–1427). Bootstrap samples that contained overlapping genome regions were thrown out. The mean SHAPE reactivities for each bootstrap sample were used to generate a frequency distribution that describes the expectation for equally sized but randomly located collections of genome elements in HIV coding regions. We obtained a P value by determining the percentage of the bootstrapped means that was lower than the mean SHAPE reactivity for the collection of genome elements. This P value is equivalent to the probability that the low SHAPE reactivity in the actual collection of genome elements occurred by chance. P values for inter-protein linkers and protein-domain junctions were 0.0482 and 0.0777, respectively. The reverse transcriptase–RNase H junction functions both as an inter-protein linker and as a protein-domain junction because it is cleaved one-half of the time. For this analysis, the reverse transcriptase–RNase H junction was counted as an inter-protein linker.

Statistical analysis of ribosome pause sites. Toeprinting data spanned 748 nucleotides (positions 670–1018 and 1243–1652; Supplementary Fig. 5). In these two reads, there were 220 nucleotides that fell within 30 nucleotides of the matrix–capsid, capsid–nucleocapsid, or nucleocapsid–p6 junctions or in the cyclophilin loop. We evaluated whether ribosomes pause preferentially near protein junctions using the binomial distribution. A total of 36 base pairs yielded toeprint signals with an intensity of 1.0 or greater. A signal of 1.0 corresponds approximately to 1.5 standard deviations above the mean; 17 of these occurred within 30 nucleotides of a protein junction. The probability of observing this distribution by chance is $P = 0.018$. This analysis was insensitive to the choice of high signal threshold. Similar P values were obtained for toeprint thresholds between 0.6 and 1.6.

Consensus structure. The *gag-pro-pol* consensus structure (Supplementary Fig. 6b) was generated by aligning the 37 reference group M HIV-1 sequences[14] using CLUSTALW[48]. Regions of covariation were identified using a sequence logo[49].

Helix energies. Helix free-energy changes (Fig. 4c, d) were calculated using the RNAstructure program[22] as the sum of the base pair stacking nearest neighbour parameters[50,51]. Duplex regions containing single nucleotide bulges were taken to be a single helix. The helix free-energy changes do not include penalties for terminal AU or GU pairs because these are, by convention in RNAstructure, associated with the loop formation free-energy changes.

RNA and protein structure display. RNA secondary structures were composed using xrna (http://rna.ucsc.edu/rnacenter/xrna); HIV protein images (Fig. 1d) were created using Visual Molecular Dynamics[52].

29. Means, R. E. *et al.* Ability of the V3 loop of simian immunodeficiency virus to serve as a target for antibody-mediated neutralization: correlation of neutralization

©2009 Macmillan Publishers Limited. All rights reserved

sensitivity, growth in macrophages, and decreased dependence on CD4. *J. Virol.* **75,** 3903–3915 (2001).
30. Graham, F. L. & van der Eb, A. J. A new technique for the assay of infectivity of human adenovirus 5 DNA. *Virology* **52,** 456–467 (1973).
31. Adachi, A. *et al.* Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol.* **59,** 284–291 (1986).
32. Chertova, E. *et al.* Envelope glycoprotein incorporation, not shedding of surface envelope glycoprotein (gp120/SU), is the primary determinant of SU content of purified human immunodeficiency virus type 1 and simian immunodeficiency virus. *J. Virol.* **76,** 5315–5325 (2002).
33. Ott, D. E. *et al.* Analysis and localization of cyclophilin A found in the virions of human immunodeficiency virus type 1 MN strain. *AIDS Res. Hum. Retroviruses* **11,** 1003–1006 (1995).
34. Thomas, J. A. *et al.* Human immunodeficiency virus type 1 nucleocapsid zinc-finger mutations cause defects in reverse transcription and integration. *Virology* **353,** 41–51 (2006).
35. Cline, A. N., Bess, J. W., Piatak, M. Jr & Lifson, J. D. Highly sensitive SIV plasma viral load assay: practical considerations, realistic performance expectations, and application to reverse engineering of vaccines for AIDS. *J. Med. Primatol.* **34,** 303–312 (2005).
36. Buckman, J. S., Bosche, W. J. & Gorelick, R. J. Human immunodeficiency virus type 1 nucleocapsid Zn^{2+} fingers are required for efficient reverse transcription, initial integration processes, and protection of newly synthesized viral DNA. *J. Virol.* **77,** 1469–1480 (2003).
37. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1,** 1610–1616 (2006).
38. Badorrek, C. S. & Weeks, K. M. Architecture of a gamma retroviral genomic RNA dimer. *Biochemistry* **45,** 12664–12672 (2006).
39. Dowell, R. D. & Eddy, S. R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5,** 71 (2004).
40. Olsen, H. S., Nelbock, P., Cochrane, A. W. & Rosen, C. A. Secondary structure is the major determinant for interaction of HIV rev protein with RNA. *Science* **247,** 845–848 (1990).
41. Wilkinson, K. A. *et al.* Influence of nucleotide identity on ribose 2′-hydroxyl reactivity in RNA. *RNA* **15,** 1314–1321 (2009).
42. Badorrek, C. S. & Weeks, K. M. RNA flexibility in the dimerization domain of a gamma retrovirus. *Nature Chem. Biol.* **1,** 104–111 (2005).
43. Knudsen, B. & Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* **31,** 3423–3428 (2003).
44. Durbin, R. & Eddy, S. *Biological Sequence Analysis: Probabalistic Models Of Proteins And Nucleic Acids* 356 (Cambridge Univ. Press, 1998).
45. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).
46. Abecasis, A. B. *et al.* Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. *J. Virol.* **81,** 8543–8551 (2007).
47. The R Development Core Team. *The R Foundation for Statistical Computing* <http://www.R-project.org> (2008).
48. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
49. Chang, T. H., Horng, J. T. & Huang, H. D. RNALogo: a new approach to display structural RNA alignment. *Nucleic Acids Res.* **36,** W91–W96 (2008).
50. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry* **37,** 14719–14735 (1998).
51. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.* **288,** 911–940 (1999).
52. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38 27–38 (1996).

©2009 **Macmillan Publishers Limited. All rights reserved**

LETTERS

A high stellar velocity dispersion for a compact massive galaxy at redshift $z = 2.186$

Pieter G. van Dokkum[1], Mariska Kriek[2] & Marijn Franx[3]

Recent studies have found that the oldest and most luminous galaxies in the early Universe are surprisingly compact[1–7], having stellar masses similar to present-day elliptical galaxies but much smaller sizes. This finding has attracted considerable attention[8–13], as it suggests that massive galaxies have grown in size by a factor of about five over the past ten billion years (10 Gyr). A key test of these results is a determination of the stellar kinematics of one of the compact galaxies: if the sizes of these objects are as extreme as has been claimed, their stars are expected to have much higher velocities than those in present-day galaxies of the same mass. Here we report a measurement of the stellar velocity dispersion of a massive compact galaxy at redshift $z = 2.186$, corresponding to a look-back time of 10.7 Gyr. The velocity dispersion is very high at 510^{+165}_{-95} km s^{-1}, consistent with the mass and compactness of the galaxy inferred from photometric data. This would indicate significant recent structural and dynamical evolution of massive galaxies over the past 10 Gyr. The uncertainty in the dispersion was determined from simulations that include the effects of noise and template mismatch. However, we cannot exclude the possibility that some subtle systematic effect may have influenced the analysis, given the low signal-to-noise ratio of our spectrum.

We observed the galaxy, dubbed 1255–0, with the Gemini Near-Infrared Spectrograph (GNIRS) on the Gemini South telescope for a total of 29 h. The de-redshifted spectrum is shown in Fig. 1a. A

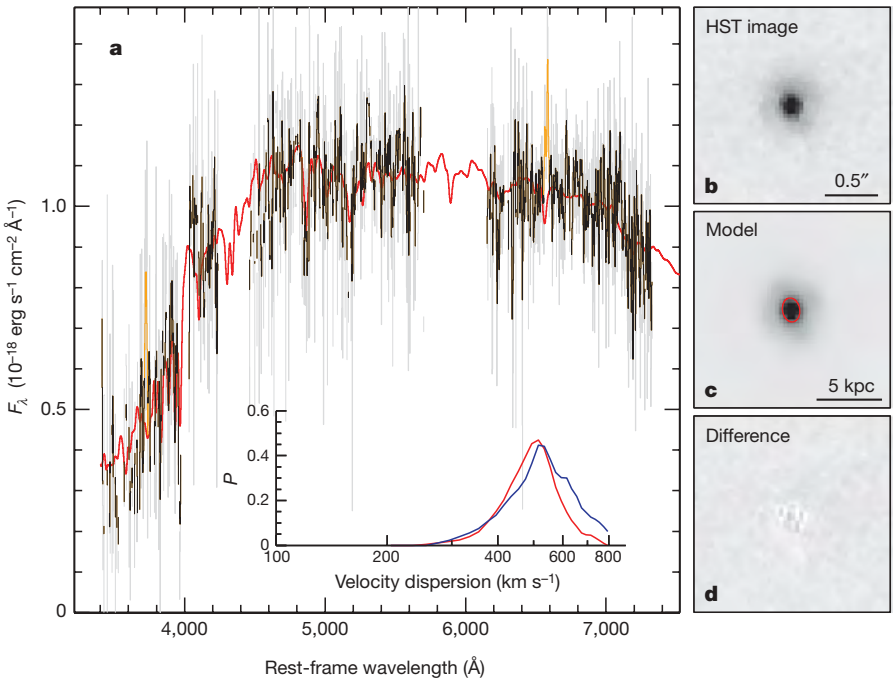


Figure 1 | Spectrum and HST images of 1255–0 at $z = 2.186$. **a**, Spectrum that was used to measure the velocity dispersion. Light grey shows the spectrum at a resolution of 5 Å (~100 km s^{-1}), which was used for the actual measurement. A smoothed version of the same data (using a 25-Å boxcar filter) is shown in black. Regions around detected emission lines are shown in orange and were excluded from the fits. The most prominent absorption lines are Hβ at wavelength 4,861 Å and Mg at 5,172 Å. The best-fitting stellar population synthesis model[14], smoothed to the best-fitting velocity dispersion, is shown in red. Inset, results of Monte Carlo simulations to determine the uncertainty in the best-fitting velocity dispersion. The curves show how often a dispersion of 510 km s^{-1} is measured given the true dispersion and noise. The two curves are for two different methods of simulating noise: shuffling the residuals of the fit in the wavelength direction (blue curve), and extracting 'empty' one-dimensional spectra from the two-dimensional spectrum (red curve). **b**–**d**, The HST NICMOS2 image of the galaxy in the H_{160} filter (**b**), the best-fitting model of the galaxy (**c**; the effective radius is indicated in red), and the residual obtained by subtracting the model from the data (**d**). The galaxy is a single, very compact object with an effective radius of 0.78 kpc. Its coordinates are right ascension $\alpha = 12$ h 54 min 59.6 s, declination $\delta = +01^\circ\ 11'\ 30''$ (J2000), its K band observed magnitude is 19.26 (Vega) and its R band observed magnitude is 24.98 (Vega)[16]. Alternative names that have been used for this object are 1256-151 (ref. 15) and 1256-0 (refs 3, 16).

[1]Astronomy Department, Yale University, 260 Whitney Avenue, New Haven, Connecticut 06511, USA. [2]Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544, USA. [3]Leiden Observatory, Leiden University, 2300 RA Leiden, The Netherlands.

©2009 Macmillan Publishers Limited. All rights reserved

detailed description of the observations and reduction, as well as an analysis of the continuum emission and detected (weak) emission lines, is presented elsewhere[14]. In ref. 14 we derive a stellar mass of $\sim 2.0 \times 10^{11} M_{\odot}$ for a Kroupa initial mass function, by fitting stellar population synthesis models to the broad band photometry and the GNIRS spectrum. The effective radius (r_e) of 1255–0 is 0.78 ± 0.17 kpc, as previously measured[3] from deep Hubble Space Telescope (HST) NICMOS2 observations. The galaxy was selected from a well-studied[3,15,16] sample of nine spectroscopically confirmed galaxies with evolved stellar populations at $z \approx 2.3$, and its properties are similar to those of other galaxies in this sample. The median stellar mass of the nine objects is $1.7 \times 10^{11} M_{\odot}$ and their median r_e is 0.9 kpc (ref. 3), a factor of $\sim$5 smaller than galaxies with similar masses at $z = 0$. The number density of these massive compact galaxies is substantial, about the same as that of galaxies in the nearby Universe that are a factor of 2–3 more massive[10].

In the present study, we use the deep Gemini spectrum to measure the stellar velocity dispersion of the galaxy, by using standard techniques for measuring the broadening of the absorption lines[17,18]. Our methodology is explained in detail in the Supplementary Information. Briefly, smoothed model spectra were fitted to the data in real space, taking observational errors into account and ignoring data with the largest uncertainties. The uncertainty in the dispersion was determined from Monte Carlo simulations of many different combinations of assumed velocity dispersions and empirical realizations of the noise. Systematic uncertainties were assessed by varying the templates (also allowing for multiple components), the masking and weighting, and the continuum filtering, and are typically much smaller than the random uncertainty. We note that the spectrum is available in electronic form (Supplementary Data).

We derive a velocity dispersion $\sigma = 510^{+165}_{-95}$ km s^{-1} for the galaxy, which is very high when compared to typical early-type galaxies in the nearby Universe. The one-sided 95% confidence lower limit is 335 km s^{-1}. Although not statistically significant, it is striking that the best-fit value exceeds the measured dispersions of all individual galaxies in the Sloan Digital Sky Survey (SDSS)[19,20]. In the SDSS, a significant fraction of galaxies with velocity dispersions in excess of 350 km s^{-1} are superpositions, which are easily identified with HST imaging[20]. As shown in Fig. 1b–d, 1255–0 is a single, nearly round object with an effective radius of $\sim$0.1 arcsec in HST images. The dispersion is also a factor of $\sim$2 higher than a previous measurement[21] from a stacked spectrum of 13 galaxies at mean redshift $\langle z \rangle = 1.6$. A direct comparison is difficult given the uncertainties associated with stacking individual spectra, but we note that the median stellar mass of the 13 galaxies is a factor of $\sim$3 smaller than that of 1255–0 and the median effective radius is a factor of 1.5 larger. The expected dispersion of these $\langle z \rangle = 1.6$ galaxies is therefore a factor of $\sim$2 lower than that of 1255–0, and the two results are consistent.

The high dispersion of 1255–0 confirms that the galaxy is very massive despite its diminutive size. The relation between dynamical mass of the galaxy, velocity dispersion and size can be expressed as $M_{\rm dyn} = C\sigma^2 r_e$, with C a constant that depends on the structure of the galaxy and other parameters. Using $\log C = 5.87$, which is the value that gives $M_{\rm dyn} \approx M_{\rm star}$ (here $M_{\rm star}$ is the stellar mass of the galaxy) for galaxies in the SDSS[6], we find $M_{\rm dyn} = 1.5^{+1.2}_{-0.5} \times 10^{11}\ M_{\odot}$. For $\log C = 6.07$, the value derived from kinematic data of present-day early-type galaxies[18], the dynamical mass is $2.4^{+1.9}_{-0.8} \times 10^{11}\ M_{\odot}$. Both estimates are in excellent agreement with the stellar mass (Fig. 2a). Put differently, the high dispersion that we measure was expected (and in fact predicted[3,10]), given our extreme size and stellar mass measurements. Quantitatively, the expected dispersion assuming $M_{\rm dyn} = M_{\rm star}$ and $5.87 \leq \log C \leq 6.07$ is in the range 470–590 km s^{-1}.

At the same time, the high dispersion confirms and extends the notion that quiescent galaxies at high redshift are structurally and dynamically very different from galaxies in the present-day Universe. Figure 2b–d shows where 1255–0 falls with respect to the relations

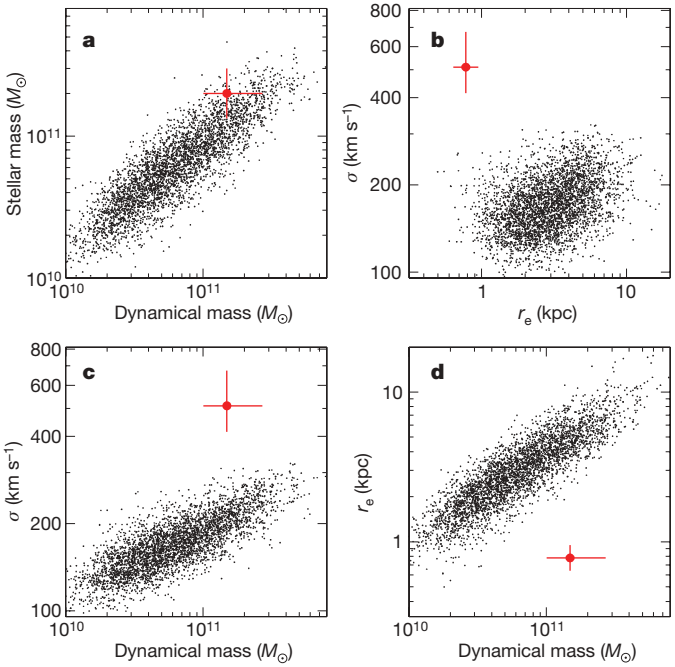


Figure 2 | Properties of 1255–0 compared to nearby galaxies. a, Relation between stellar mass and dynamical mass. Small symbols are galaxies in the SDSS[6] in the redshift range 0.05–0.07, and the large red symbol is galaxy 1255–0 at $z = 2.186$. Our definition of dynamical mass, $\log M_{\rm dyn} = 5.87 + 2\log\sigma + \log r_e$, leads to a one-to-one correspondence between stellar mass and dynamical mass for SDSS galaxies. Despite its small size 1255–0 has a very high mass, similar to elliptical galaxies today. The dynamical mass is consistent (within 1 s.d.) with the stellar mass that was estimated[14] from fitting stellar population synthesis models to the photometry. **b–d**, Relations between velocity dispersion, effective radius and dynamical mass. Note that these three panels do not depend on stellar populations (except indirectly through the fact that the spectrum and the HST image are weighted by luminosity, not mass). It is clear that the structure and kinematics of 1255–0 are fundamentally different from those of nearby galaxies, and significant evolution is required to bring this object to the local relations. Error bars, 1 s.d.

between velocity dispersion, size and dynamical mass defined by SDSS galaxies. The galaxy is offset from the local relations, consistent with previous studies that were based on stellar masses derived from photometric data[3,6]. At fixed dynamical mass, the dispersion is higher by a factor of $\sim$2.5 and the effective radius is smaller by a factor of $\sim$6. The most dramatic offset is in the $\log\sigma$–$\log r_e$ plane (Fig. 2b). These two quantities are measured directly and independently, and (to first order) do not depend on stellar populations.

The extreme compactness of massive high-redshift galaxies is qualitatively consistent with models in which the central parts of massive galaxies form early in highly dissipative processes[22], although it remains to be seen whether such models can produce objects with the size and velocity dispersion of 1255–0. In particular, it may be difficult to funnel gas clumps into an extremely compact configuration without forming stars at larger radii. Regardless of the details of the model, in its star-forming phase at $z \gtrsim 3.5$ the galaxy probably had a very compact molecular gas distribution with a rotation velocity of $\sim$700 km s^{-1}. The median rotation velocity ($V_{\rm rot}$) of CO in submillimetre galaxies at $z = 2$–3.5 has been found[23] to be $\sim$470 km s^{-1} (assuming $V_{\rm rot} = 0.6 \times$ FWHM, where FWHM is full-width at half-maximum); this is a high value by most standards, but still somewhat lower than what we expect for the progenitors of galaxies such as 1255–0. There is not yet much information on the gas dynamics of massive galaxies at redshifts $z > 3.5$. The $z = 6.4$ quasar SDSS J1148+5251 has a relatively small CO linewidth of $V_{\rm rot} \approx 170$ km s^{-1} (ref. 24), but it may be that quasars are biased low because their gas disks are preferentially seen face-on[25]. It is

©2009 Macmillan Publishers Limited. All rights reserved

obviously not clear whether the molecular gas in the progenitor of 1255-0 was ever in a regular disk; it would be interesting to determine whether 1255–0 shows rotation, but that requires imaging (or spectroscopy) of higher spatial resolution than is currently available.

A problem that is perhaps even more vexing than the origin of galaxies such as 1255–0 is their subsequent evolution onto the local relations between size, velocity dispersion and mass. The simplest explanation is that the mass and/or size measurements of the compact galaxies are incorrect[3,13], but this is difficult to maintain given the dynamical measurement presented here. We are left with the conclusion that very significant structural and dynamical changes are required to bring massive, quiescent high-redshift galaxies to the local relations. This cannot easily be achieved through star formation, as the compact high-redshift galaxies already appear to have stopped forming new stars, consistent with the old ages inferred for the stars in today's most massive galaxies. Among the models that have been proposed[8–13], minor mergers may be the most effective single mechanism, as simple virial arguments suggest that the velocity dispersion changes by a factor of $f_r^{-1/4}$ for a factor of f_r change in radius[10,11]. However, it is an open question whether mergers alone can 'puff up' galaxies by the required amount, as the precise effect depends on the accretion rate, the masses, orbits and gas content of accreted galaxies, angular momentum transfer between stars and dark matter, and on possible evolution in the contribution of dark matter to the measured kinematics[26]. Finally, we note that evolution in the velocity dispersion of galaxies implies evolution in the black hole mass–σ relation[27,28], such that black hole masses must be lower at fixed σ at high redshift.

While confirming that the velocity dispersions of compact galaxies are high, our measurement is obviously not sufficiently accurate to properly characterize the evolution of the relations in Fig. 2. A 1 s.d. error of 25% in the velocity dispersion implies an error of 56% in the dynamical mass, and further progress requires dispersions with uncertainties $\lesssim$10% for much larger samples. New spectrographs being readied for use on 8-m-class telescopes, combined with new wide-field imaging surveys that can provide sufficiently bright targets, are expected to revolutionize this field in the next few years. As indicated here, such observations are crucial for calibrating stellar masses at high redshift, and for measuring the structural and dynamical evolution of massive galaxies from the time that their star formation was quenched to the present.

Received 20 April; accepted 12 June 2009.

1. Trujillo, I. *et al.* The size evolution of galaxies since $z \sim 3$: combining SDSS, GEMS, and FIRES. *Astrophys. J.* **650,** 18–41 (2006).
2. Toft, S. *et al.* Hubble Space Telescope and Spitzer imaging of red and blue galaxies at $z \sim 2.5$: a correlation between size and star formation activity from compact quiescent galaxies to extended star-forming galaxies. *Astrophys. J.* **671,** 285–302 (2007).
3. van Dokkum, P. G. *et al.* Confirmation of the remarkable compactness of massive quiescent galaxies at $z \sim 2.3$: early-type galaxies did not form in a simple monolithic collapse. *Astrophys. J.* **677,** L5–L8 (2008).
4. Cimatti, A. *et al.* GMASS ultradeep spectroscopy of galaxies at $z \sim 2$. II. Superdense passive galaxies: how did they form and evolve? *Astron. Astrophys.* **482,** 21–35 (2008).
5. van der Wel, A. *et al.* Recent structural evolution of early-type galaxies: size growth from $z = 1$ to $z = 0$. *Astrophys. J.* **688,** 48–58 (2008).
6. Franx, M. *et al.* Structure and star formation in galaxies out to $z = 3$: evidence for surface density dependent evolution and upsizing. *Astrophys. J.* **688,** 770–788 (2008).
7. Damjanov, I. *et al.* Red nuggets at $z \sim 1.5$: compact passive galaxies and the formation of the Kormendy relation. *Astrophys. J.* (in the press); preprint at ⟨http://arXiv.org/abs/0807.1744⟩ (2008).
8. Naab, T., Johansson, P. H., Ostriker, J. P. & Efstathiou, G. Formation of early-type galaxies from cosmological initial conditions. *Astrophys. J.* **658,** 710–720 (2008).
9. Fan, L., Lapi, A., De Zotti, G. & Danese, L. The dramatic size evolution of elliptical galaxies and the quasar feedback. *Astrophys. J.* **689,** L101–L104 (2008).
10. Bezanson, R. *et al.* The relation between compact, quiescent high redshift galaxies and massive nearby elliptical galaxies: evidence for hierarchical, inside-out growth. *Astrophys. J.* (in the press); preprint at ⟨http://arXiv.org/abs/0903.2044⟩ (2009).
11. Naab, T., Johansson, P. H. & Ostriker, J. P. Minor mergers and the size evolution of elliptical galaxies. *Astrophys. J.* (submitted); preprint at ⟨http://arXiv.org/abs/0903.1636⟩ (2009).
12. van der Wel, A., Bell, E. F., van den Bosch, F. C., Gallazzi, A. & Rix, H.-W. On the size and co-moving mass density evolution of early-type galaxies. *Astrophys. J.* (submitted); preprint at ⟨http://arXiv.org/abs/0903.4857⟩ (2009).
13. Hopkins, P. F. *et al.* Compact high-redshift galaxies are the core of present-day massive spheroids. *Astrophys. J.* (submitted); preprint at ⟨http://arXiv.org/abs/0903.2479⟩ (2009).
14. Kriek, M. *et al.* An ultra-deep near-infrared spectrum of a compact quiescent galaxy at $z = 2.2$. *Astrophys. J.* (in the press); preprint at ⟨http://arXiv.org/abs/0905.1692⟩ (2009).
15. Kriek, M. *et al.* Spectroscopic identification of massive galaxies at $z \sim 2.3$ with strongly suppressed star formation. *Astrophys. J.* **649,** L71–L74 (2006).
16. Kriek, M. *et al.* A near-infrared spectroscopic survey of *K*-selected galaxies at $z \sim 2.3$: redshifts and implications for broadband photometric studies. *Astrophys. J.* **677,** 219–237 (2008).
17. Franx, M., Illingworth, G. & Heckman, T. Major and minor axis kinematics of 22 ellipticals. *Astrophys. J.* **344,** 613–636 (1989).
18. van Dokkum, P. G. & Stanford, S. A. The fundamental plane at $z = 1.27$: first calibration of the mass scale of red galaxies at redshifts $z > 1$. *Astrophys. J.* **585,** 78–89 (2003).
19. Bernardi, M. *et al.* A search for the most massive galaxies: double trouble? *Astron. J.* **391,** 1191–1199 (2006).
20. Bernardi, M. *et al.* A search for the most massive galaxies – II. Structure, environment, and formation. *Mon. Not. R. Astron. Soc.* **391,** 1191–1199 (2008).
21. Cenarro, A. & Trujillo, I. Mild velocity dispersion evolution of spheroid-like massive galaxies since $z \sim 2$. *Astrophys. J.* **696,** L43–L46 (2009).
22. Dekel, A. *et al.* Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* **457,** 451–454 (2009).
23. Greve, T. R. *et al.* An interferometric CO survey of luminous submillimetre galaxies. *Mon. Not. R. Astron. Soc.* **359,** 1165–1183 (2005).
24. Bertoldi, F. *et al.* High-excitation CO in a quasar host galaxy at $z = 6.42$. *Astron. Astrophys.* **409,** L47–L50 (2003).
25. Narayanan, D. *et al.* The nature of CO emission from $z \sim 6$ quasars. *Astrophys. J.* **174** (Supp.), 13–30 (2008).
26. Boylan-Kolchin, M., Ma, C.-P. & Quataert, E. Red mergers and the assembly of massive elliptical galaxies: the fundamental plane and its projections. *Mon. Not. R. Astron. Soc.* **369,** 1081–1089 (2006).
27. Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539,** L9–L12 (2000).
28. Gebhardt, K. *et al.* A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J.* **539,** L13–L16 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This Letter is based on observations obtained at the Gemini Observatory and with the HST. This work was supported by NASA and the NSF. We thank I. Labbé, G. Illingworth, D. Marchesini and R. Quadri for their contributions in the initial stages of this project.

Author Contributions P.G.v.D. wrote the Gemini proposal, did the observations, measured the velocity dispersion, wrote the Letter and led the interpretation. M.K. reduced the Gemini spectrum, determined the stellar mass and contributed to the interpretation. M.F. independently measured the velocity dispersion and contributed to the analysis and interpretation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.G.v.D. (pieter.vandokkum@yale.edu).

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

Observed variations of methane on Mars unexplained by known atmospheric chemistry and physics

Franck Lefèvre[1] & François Forget[2]

The detection of methane on Mars[1–3] has revived the possibility of past or extant life on this planet, despite the fact that an abiogenic origin is thought to be equally plausible[4]. An intriguing aspect of the recent observations of methane on Mars is that methane concentrations appear to be locally enhanced and change with the seasons[3]. However, methane has a photochemical lifetime of several centuries, and is therefore expected to have a spatially uniform distribution on the planet[5]. Here we use a global climate model of Mars with coupled chemistry[6–8] to examine the implications of the recently observed variations of Martian methane for our understanding of the chemistry of methane. We find that photochemistry as currently understood does not produce measurable variations in methane concentrations, even in the case of a current, local and episodic methane release. In contrast, we find that the condensation–sublimation cycle of Mars' carbon dioxide atmosphere can generate large-scale methane variations differing from those observed. In order to reproduce local methane enhancements similar to those recently reported[3], we show that an atmospheric lifetime of less than 200 days is necessary, even if a local source of methane is only active around the time of the observation itself. This implies an unidentified methane loss process that is 600 times faster than predicted by standard photochemistry. The existence of such a fast loss in the Martian atmosphere is difficult to reconcile with the observed distribution of other trace gas species. In the case of a destruction mechanism only active at the surface of Mars, destruction of methane must occur with an even shorter timescale of the order of ~1 hour to explain the observations. If recent observations of spatial and temporal variations of methane are confirmed, this would suggest an extraordinarily harsh environment for the survival of organics on the planet.

The fact that methane concentration varies with time and location on Mars contradicts the logic that a gas of lifetime much longer than the time required for global mixing should have a constant and spatially uniform distribution. To examine the implications of the recently observed variations of Martian methane, we implemented methane chemistry in the Laboratoire de Météorologie Dynamique (LMD) global climate model (GCM) of Mars with on-line photochemistry[6–8]. In the 'conventional' atmospheric chemistry scheme, which explains correctly the observed distribution of methane on Earth, loss of methane on Mars occurs primarily by photolysis at heights above 60 km, and by oxidation by OH and $O(^1D)$ at lower altitudes. These constituents are produced respectively by the photolysis of water and ozone. The fact that the LMD GCM reproduces closely the observed seasonal and geographical variations of these species[8] is an important prerequisite for a precise estimate of the fate of methane in the Martian atmosphere.

To determine the atmospheric lifetime of methane, we first initialized the GCM with a uniform mixing ratio and monitored the exponential decay of methane in a long-term simulation that did not include any source. We find that the global atmospheric mass of methane is reduced by a factor of e after 330 terrestrial years. This lifetime is consistent with past estimations based on globally averaged models (250–670 terrestrial years[2,9,10]), but here integrates the effects of spatial and seasonal variations in ultraviolet flux, water vapour and ozone. From this estimation, the source flux of methane at the Martian surface must be 260 tonnes per year ($260\,t\,yr^{-1}$) for a steady-state value of 10 p.p.b.v. (parts per billion by volume). This may be compared with the terrestrial value of $582 \times 10^6\,t\,yr^{-1}$ (ref. 11).

Can such a faint source create variations in the observed methane field? Evidently, such variations are favoured if the source itself shows some degree of spatial or temporal variability. To investigate this possibility, we introduced a highly localized and sporadic source in the GCM. We chose the area and timing of the methane release to coincide with the important local maximum (40–50 p.p.b.v.) observed[3] in northern summer 2003: methane is released at the surface in a single grid cell of the model located in Syrtis Major (10° N, 50° E), and the emission is assumed to occur for only 60 sols (one sol is a Martian day) around solar longitude $L_s = 150°$. The amount of methane injected into the atmosphere during this period is constrained to balance the global photochemical loss integrated over the Martian year. Figure 1a displays the methane distribution obtained during the period of emission. The local release of methane does not produce any significant enhancement or plume in the source region. In contrast with the observation, the model shows an essentially well-mixed distribution over most of the planet. A striking feature, however, is the large methane enrichment that results from the condensation of CO_2 gas at high southern latitudes. The enrichment in non-condensable species during the formation of the seasonal CO_2 ice cap is a well established process, derived from observations of argon by the Gamma Ray Spectrometer on board Mars Odyssey[12,13]. In relative terms, the enhancement (or depletion) factor due to CO_2 condensation (or sublimation) should be identical for all non-condensable species, and hence for methane and argon. The predicted enrichment factor of 4 to 5 for methane at $L_s = 150°$ is in good quantitative agreement with the argon enhancement measured by the Gamma Ray Spectrometer (ref. 14; Supplementary Information).

Moving the source to other regions has no effect on the GCM methane field. For instance, we attempted to reproduce the methane maximum reported in southern spring/summer over the south of Tharsis[15]. The signature of the short-lived release at $L_s = 270°$ remains invisible in the mixing ratio map (Fig. 1b). At the same time, the sublimation of the seasonal CO_2 ice cap in the model leads to a depletion of methane that reaches ~70% at high southern latitudes, identical to the depletion of carbon monoxide (another non-condensable species) observed by the Mars Reconnaissance Orbiter[16].

Figure 1c shows that the condensation–sublimation cycle of CO_2 drives the entire seasonal cycle of methane calculated by the GCM

[1]LATMOS, [2]Laboratoire de Météorologie Dynamique, UPMC Université Paris 06, CNRS, Paris 75005, France.

©2009 Macmillan Publishers Limited. All rights reserved

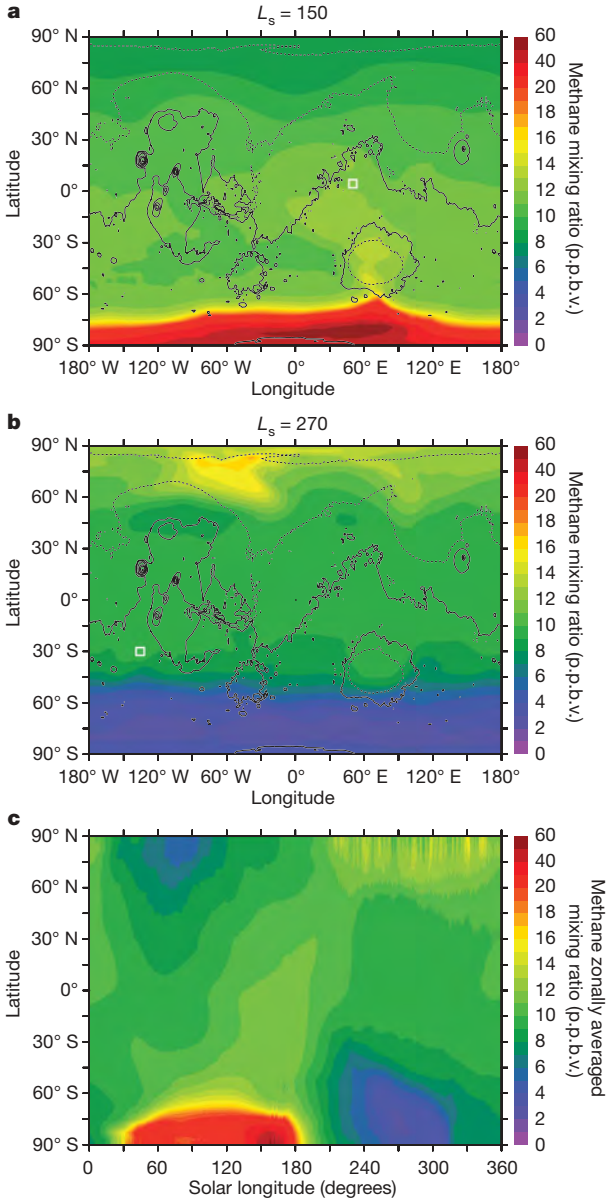


Figure 1 | Column-averaged methane mixing ratio calculated by the global climate–chemical model. The simulations include a local source at the surface indicated by the white square. The amount of methane released from the source balances the global photochemical loss integrated over the Martian year, assuming an equilibrium value of 10 p.p.b.v. **a**, Methane is emitted at 10° N 50° E for a period of 60 sols centred on $L_s = 150°$. **b**, Methane is emitted at 30° S 135° W for a period of 60 sols centred on $L_s = 270°$. Although shown during the period of emission, in both cases the methane maps do not show any signature of the local source. **c**, Seasonal evolution of the zonally averaged mixing ratio for the experiment shown in **a**.

with standard photochemistry. The largest variations are found during southern spring ($L_s = 180$–$270°$), when non-condensable species decrease rapidly south of 75° S. At mid-latitudes, the GCM simulation shows a weak seasonal modulation (2–3 p.p.b.v.) of the methane mixing ratio that is mainly related to slow variation in the global pressure of CO_2. This contrasts with observations by the Planetary Fourier Spectrometer on board Mars Express, which suggest a methane increase from 5 p.p.b.v. to 20 p.p.b.v. between $L_s = 360°$ and $L_s = 50°$ (ref. 17). In addition, the GCM does not show any methane diurnal cycle, and no correlation is found with water vapour, contrary to what seems to be observed from Planetary Fourier Spectrometer spectra[17].

Our results demonstrate that even in the most favourable case of a highly localized source only active around the time of the observation

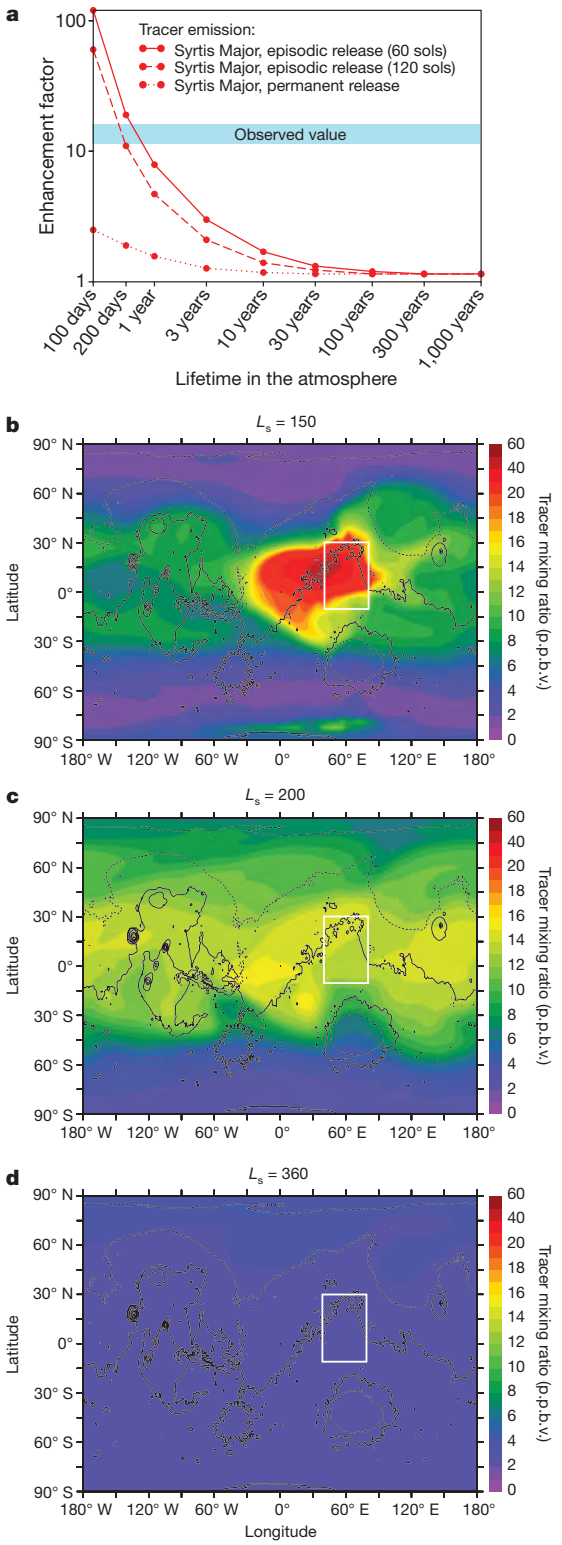


Figure 2 | Idealized tracer experiments. **a**, Maximum enhancement created by a local source of tracer in a column-averaged tracer field, as a function of tracer lifetime in the Martian atmosphere. The release of tracer is either permanent or episodic (60 or 120 sols, centred on $L_s = 150°$) from the region where enhanced methane was observed[3]. Tracer loss occurs at all altitudes according to the designated lifetime. The enhancement factor is defined as the ratio of the tracer abundance in the emission area to the homogeneously mixed value at vernal equinox ($L_s = 360°$), and has an observed value of ~12 (ref. 3). **b**, Column-averaged mixing ratio at $L_s = 150°$ of a tracer with an atmospheric lifetime of 200 terrestrial days emitted for 120 sols ($L_s = 120$–$183°$). The area of emission is indicated by the white rectangle. **c**, As **b** but at $L_s = 200°$. **d**, As **b** but at $L_s = 360°$.

©2009 Macmillan Publishers Limited. All rights reserved

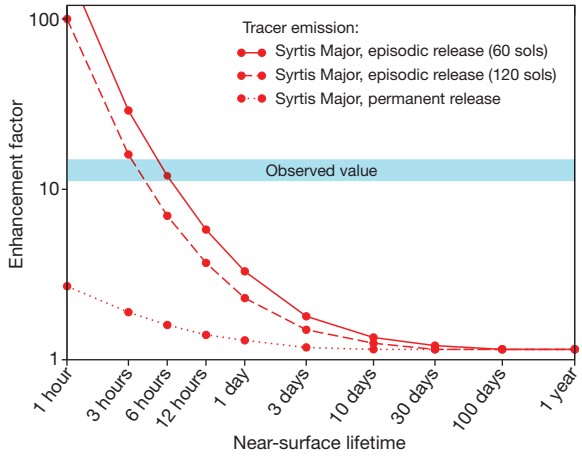


Figure 3 | Maximum enhancement created by a local source of tracer in a column-averaged tracer field, as a function of tracer lifetime at the surface of Mars. Scenarios of emission are identical to those of Fig. 2a. Tracer loss here only occurs in a 10-m-thick layer in contact with the surface, according to the designated lifetime.

itself, methane chemistry as currently understood cannot explain the spatial and temporal enhancements that have been reported. To create variations other than those due to CO_2 condensation, a considerably more intense source is required. This in turn implies a greater sink, and hence a shorter lifetime in order to maintain the same quantity of methane in the atmosphere. To determine this lifetime in our simulations, we released idealized tracers from the region of Syrtis Major where enhanced methane was observed in 2003 (ref. 3). A particular atmospheric lifetime was attributed to each tracer, which we assumed to be identical on all vertical levels. For each emission scenario (permanent or restricted to 60 or 120 sols), the mass of tracer injected into the atmosphere balances exactly the integrated loss over the Martian year. Figure 2a quantifies the maximum tracer enhancement obtained in Syrtis Major for lifetimes between 100 terrestrial days and 100 terrestrial years. The enhancement factor is defined as the ratio of the tracer mixing ratio in the emission area to the homogeneously distributed value at vernal equinox ($L_s = 0°$), and has an observed value of ~12 (ref. 3). This value is reached in our simulations for an episodic release and if the gas has a lifetime of about 200 terrestrial days, in agreement with the value obtained in ref. 3. Under these conditions, the GCM reproduces closely the observed spatial and temporal variability of the methane distribution. During the period of emission, at $L_s = 150°$, the intense release of tracer maintains a plume of strong values (>40 p.p.b.v.) over Syrtis Major (Fig. 2b). Rapid dispersion by the atmospheric circulation then occurs. At $L_s = 200°$ (Fig. 2c), less than 30 sols after the source ceased to emit, the region of emission is no longer identifiable in the tracer map. Mixing combined with the reduced chemical lifetime eventually leads to a quasi-uniform mixing ratio of 2–3 p.p.b.v. at $L_s = 360°$ (Fig. 2d). This optimum quantitative agreement with the methane observations is obtained with ~150,000 t of methane emitted by the sporadic source. This amount is comparable to the yearly geochemical production of methane by serpentinization (50,000–130,000 t yr^{-1}) along the entire Mid-Atlantic Ridge on Earth (ref. 18, and R. Keir, personal communication).

A lifetime of 200 terrestrial days implies the existence of an unknown methane sink that is 600 times more efficient than the loss predicted by the current consensus on terrestrial atmospheric chemistry. It has been proposed that methane can be destroyed on Mars by electrochemical processes triggered by the strong electric fields generated during dust storms[4,19–21]. We investigated this hypothesis by implementing in the GCM the dissociation of methane, CO_2 and H_2O by energized electrons (refs 19, 20; see also Methods). In dust storms, these processes are expected to increase the methane destruction rate and to produce vast amounts of hydrogen peroxide, H_2O_2 (refs 4, 20, 21). The H_2O_2 mixing ratio was determined to be 18 p.p.b.v. at 20° S in the dust storm season ($L_s = 250°$, equivalent dust visible opacity of ~1 at 7 hPa) that followed the detection of the methane plume in 2003 (ref. 22). This amount of H_2O_2 is well reproduced by the GCM without the need for electrically charged dust[8], which provides a lower limit on the dust threshold at which a large-scale impact of electrochemical processes on the chemistry can be envisaged. Using this constraint, we tested various parameterizations of the electric field coupled to the seasonal evolution of dust opacity observed by the Thermal Emission Spectrometer[23] in 2002–04. To obtain an electrochemical loss of methane that approaches the large-scale methane decrease observed between $L_s \approx 150°$ and $L_s \approx 360°$, we find that the electric field must be close to the breakdown field strength value (~25 kV m^{-1}) in all the regions with visible dust opacity of ~2 or above. The possibility that such extreme bulk electric fields can be sustained in the Martian lower atmosphere has recently been severely questioned[24]. Furthermore, the electrochemical dissociation of CO_2 calculated in the same conditions rapidly leads to unrealistically large amounts of CO in the model (>15,000 p.p.m.v. at $L_s = 360°$), exceeding the observations by a factor of ~20 (ref. 16).

Alternatively, destruction of methane could take place in the Martian regolith. Heterogeneous loss of methane has been shown to be slow on mineral surfaces analogous to Martian materials[25], but the presence of one or more strong oxidants in the Martian soil could accelerate this process. H_2O_2 is a good candidate, which could form through the interaction of minerals with water[26,27] or accumulate in the soil following the precipitation of condensed H_2O_2 produced in dust storms or dust devils[21].

To simulate the loss of methane in the regolith, we assumed that the idealized tracers used previously are only destroyed in the first 10-m-thick atmospheric layer in contact with the surface. Clearly, this hypothesis places an even greater burden on the efficiency of the methane loss process. Figure 3 shows that the global-scale lifetime of methane in this near-surface layer must be in the range 3–6 h to obtain an enhancement factor that matches the observation. Given the turbulent fluxes calculated in the layer, this implies a lifetime of less than 1 h at the atmosphere–regolith interface. Such a lifetime suggests that organics are quite readily scavenged from the modern Martian environment, if reactions in the surface are the only cause of the observed methane variations. This would leave little hope that life as we know it can exist at present or that evidence of past life can be preserved in the shallow surface layer. The next Mars Science Laboratory (2011) and Exomars (2016) rovers will allow examination of this hypothesis, and should be able to determine *in situ* whether methane variations exist on Mars.

METHODS SUMMARY

The LMD GCM was integrated at a resolution of 3.75° latitude × 5.625° longitude, on 32 vertical levels from the ground up to about 120 km. The photochemical code implemented in the GCM is an evolution of the model extensively described previously[7], with updated kinetics and photochemical data[28].

The simulations including electrochemical processes use the production rates of CO/O^- and OH/H^- pairs calculated as a function of the ambient electric field E in ref. 19. These rates published for surface conditions are scaled in the GCM by the local densities of CO_2 and H_2O. For each value of the electric field, the production of OH/H^- is therefore constrained by the local vertical distribution of water vapour. The electron dissociation rate of methane as a function of E is taken from ref. 20.

The efficiency of electrochemical processes grows exponentially with E up to the atmospheric breakdown level, estimated to be ~25 kV m^{-1} (ref. 29). However, the actual value of E in Martian dust storms has never been measured. To get around this difficulty, we used the dust opacity τ measured by the Thermal Emission Spectrometer[23] as a proxy for dust storm activity, and explored the sensitivity of our results to various linear or nonlinear relationships between τ and E. The τ field prescribed in our experiments[30] is three-dimensional, evolves with time, and reproduces the Thermal Emission Spectrometer observations of Martian year 26 (April 2002–March 2004), characterized by a peak in dust storm activity at $L_s = 315°$ (December 2003). This approach has the advantages of

©2009 Macmillan Publishers Limited. All rights reserved

providing a realistic distribution of the large-scale regions where strong E fields may be expected, and of constraining the GCM with observations of dust performed in the same Martian year during which enhanced methane was identified[3] (January–March 2003) and H_2O_2 was measured[22] (September 2003).

Received 18 January; accepted 12 June 2009.

1. Formisano, V., Atreya, S. K., Encrenaz, T., Ignatiev, N. & Giuranna, M. Detection of methane in the atmosphere of Mars. *Science* **306,** 1758–1761 (2004).
2. Krasnopolsky, V. A., Maillard, J. P. & Owen, T. C. Detection of methane in the martian atmosphere: evidence for life? *Icarus* **172,** 537–547 (2004).
3. Mumma, M. *et al.* Strong release of methane on Mars in northern summer 2003. *Science* **323,** 1041–1045 (2009).
4. Atreya, S. K., Mahaffy, P. R. & Wong, A. S. Methane and related trace species on Mars: origin, loss, implications for life, and habitability. *Planet. Space Sci.* **55,** 358–369 (2007).
5. Krasnopolsky, V. A. Some problems related to the origin of methane on Mars. *Icarus* **180,** 359–367 (2006).
6. Forget, F. *et al.* Improved general circulation models of the Martian atmosphere from the surface to above 80 km. *J. Geophys. Res.* **104,** 24155–24176 (1999).
7. Lefèvre, F., Lebonnois, S., Montmessin, F. & Forget, F. Three-dimensional modeling of ozone on Mars. *J. Geophys. Res.* **109,** doi:10.1029/2004JE002268 (2004).
8. Lefèvre, F. *et al.* Heterogeneous chemistry in the atmosphere of Mars. *Nature* **454,** 971–975 (2008).
9. Summers, M. E., Lieb, B. J., Chapman, E. & Yung, Y. L. Atmospheric biomarkers of subsurface life on Mars. *Geophys. Res. Lett.* **29,** doi:10.1029/2002GL015377 (2002).
10. Wong, A. S., Atreya, S. K. & Encrenaz, T. Chemical markers of possible hot spots on Mars. *J. Geophys. Res.* **108,** doi:10.1029/2002JE002003 (2003).
11. Solomon, S, *et al.* (eds) *Climate Change 2007: The Physical Science Basis* (Cambridge Univ. Press, 2007).
12. Sprague, A. L. *et al.* Mars' south polar Ar enhancement: a tracer for south polar meridional mixing. *Science* **306,** 1364–1367 (2004).
13. Sprague, A. L. *et al.* Mars' atmospheric argon: tracer for understanding Martian atmospheric circulation and dynamics. *J. Geophys. Res.* **112,** doi:10.1029/2005JE002597 (2007).
14. Forget, F., Millour, E., Montabone, L. & Lefèvre, F. Non-condensable gas enrichment and depletion in the martian polar regions. Presented at *Third Workshop on Mars Modeling and Observations* ⟨http://www.lpi.usra.edu/meetings/modeling2008/pdf/9106.pdf⟩ (2008).
15. Mumma, M. *et al.* Absolute measurements of methane on Mars: the current status. Presented at *Third Workshop on Mars Modeling and Observations* ⟨http://www.lpi.usra.edu/meetings/modeling2008/pdf/9099.pdf⟩ (2008).
16. Smith, M. D., Wolff, M. J., Clancy, R. T. & Murchie, S. L. Compact Reconnaissance Imaging Spectrometer observations of water vapor and carbon monoxide. *J. Geophys. Res.* **114,** doi:10.1029/2008JE003288 (2009).
17. Geminale, A., Formisano, V. & Giuranna, M. Methane in Martian atmosphere: average spatial, diurnal, and seasonal behaviour. *Planet. Space Sci.* **56,** 1194–1203 (2008).
18. Keir, R. S. *et al.* Methane and methane carbon isotope ratios in the Northeast Atlantic including the Mid-Atlantic Ridge (50°N). *Deep-Sea Res. I* **52,** 1043–1070 (2005).
19. Delory, G. T. *et al.* Oxidant enhancement in martian dust devils and storms: storm electric fields and electron attachment. *Astrobiology* **6,** 451–462 (2006).
20. Farrell, W. M., Delory, G. T. & Atreya, S. K. Martian dust storms as a possible sink of atmospheric methane. *Geophys. Res. Lett.* **33,** doi:10.1029/2006GL027210 (2006).
21. Atreya, S. K. *et al.* Oxidant enhancement in Martian dust devils and storms: implications for life and habitability. *Astrobiology* **6,** 439–450 (2006).
22. Clancy, R. T., Sandor, B. J. & Moriarty-Schieven, G. H. A measurement of the 362 GHz absorption line of Mars atmospheric H_2O_2. *Icarus* **168,** 116–121 (2004).
23. Smith, M. D. Interannual variability in TES atmospheric observations of Mars during 1999–2003. *Icarus* **167,** 148–165 (2004).
24. Kok, J. F. & Renno, N. O. Electrification of wind-blown sand on Mars and its implications for atmospheric chemistry. *Geophys. Res. Lett.* **36,** doi:10.1029/2008GL036691 (2009).
25. Gough, R. V., Tolbert, M. A., McKay, C. P. & Toon, O. B. Methane adsorption on Martian soil analogs: a possible abiogenic explanation for methane variability. Presented at *40th Lunar and Planetary Science Conference* ⟨http://www.lpi.usra.edu/meetings/lpsc2009/pdf/1968.pdf⟩ (2009).
26. Hurowitz, J. A., Tosca, N. J., McLennan, S. M. & Schoonen, M. A. A. Production of hydrogen peroxide in Martian and lunar soils. *Earth Planet. Sci. Lett.* **255,** 41–52 (2007).
27. Davila, A. F. *et al.* Subsurface formation of oxidants on Mars and implications for the preservation of organic biosignatures. *Earth Planet. Sci. Lett.* **272,** 456–463 (2008).
28. Sander, S. P. *et al. Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies, Evaluation Number 15* (JPL Publication 06-2, Jet Propulsion Laboratory, 2006).
29. Melnik, O. & Parrot, M. Electrostatic discharge in Martian dust storms. *J. Geophys. Res.* **103,** 29107–29117 (1998).
30. Montabone, L., Lewis, S. R. & Read, P. L. Interannual variability of Martian dust storms in assimilation of several years of Mars global surveyor observations. *Adv. Space Res.* **36,** 2146–2155 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The LMD Martian global climate model has been developed with the support of CNRS, ESA and CNES. We thank R. M. Haberle and F. Montmessin for their contributions to an early phase of this work, as well as P.-Y. Meslin and R. Wordsworth for discussions.

Author Contributions F. L. and F. F. conceived the experiments and wrote the paper. F. L. performed the experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.L. (franck.lefevre@upmc.fr).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 460|6 August 2009|doi:10.1038/nature08171

LETTERS

Observation of strong coupling between a micromechanical resonator and an optical cavity field

Simon Gröblacher[1,2], Klemens Hammerer[3,4], Michael R. Vanner[1,2] & Markus Aspelmeyer[1]

Achieving coherent quantum control over massive mechanical resonators is a current research goal. Nano- and micromechanical devices can be coupled to a variety of systems, for example to single electrons by electrostatic[1,2] or magnetic coupling[3,4], and to photons by radiation pressure[5–9] or optical dipole forces[10,11]. So far, all such experiments have operated in a regime of weak coupling, in which reversible energy exchange between the mechanical device and its coupled partner is suppressed by fast decoherence of the individual systems to their local environments. Controlled quantum experiments are in principle not possible in such a regime, but instead require strong coupling. So far, this has been demonstrated only between microscopic quantum systems, such as atoms and photons (in the context of cavity quantum electrodynamics[12]) or solid state qubits and photons[13,14]. Strong coupling is an essential requirement for the preparation of mechanical quantum states, such as squeezed or entangled states[15–18], and also for using mechanical resonators in the context of quantum information processing, for example, as quantum transducers. Here we report the observation of optomechanical normal mode splitting[19,20], which provides unambiguous evidence for strong coupling of cavity photons to a mechanical resonator. This paves the way towards full quantum optical control of nano- and micromechanical devices.

A common feature of all coupled quantum systems is that their dynamics are dominated by the competition between the joint coupling rate and the rates at which the coupled systems decohere into their local environments. Only for sufficiently strong coupling can the effects of decoherence be overcome. This so-called 'strong coupling regime' is, in all cases, indispensable for the experimental investigation of a manifold of quantum phenomena. Nano- and micro-optomechanical oscillators are currently emerging as a new 'textbook' example for coupled quantum systems. In this case, a single electromagnetic field mode is coupled to a (nano- or micrometre sized) mechanical oscillator. In analogy to cavity quantum electrodynamics (cQED), one can identify strong coupling as the regime where the coupling rate g exceeds both the cavity amplitude decay rate κ and the mechanical damping rate γ_m—as required, for example, in refs 15–17. Another class of proposals requires the weaker condition of 'large cooperativity', that is, $g > \sqrt{\kappa\gamma_m}$ (refs 18, 21). Strong coupling, ideally in combination with the preparation of zero entropy initial states (for example, by ground-state cooling of the mechanical resonator), is essential to obtain (quantum) control over this new domain of quantum physics. Whereas ground state preparation is a goal of continuing research (in which much progress has been made, in particular by using optical laser cooling techniques[22]), here we demonstrate strong optomechanical coupling using state-of-the-art micromechanical resonators.

Consider the canonical situation in which a mechanical resonator is coupled to the electromagnetic field of a high-finesse cavity via momentum transfer of the cavity photons (Fig. 1). The system naturally comprises two coupled oscillators: the electromagnetic field at cavity frequency ω_c (typically of the order of 10^{15} Hz) and the mechanical resonator at frequency ω_m ($\sim 10^7$ Hz). At first sight, the large discrepancy in the oscillator frequencies seems to inhibit any coupling; it is, however, alleviated by the fact that the cavity is driven by a laser field at frequency ω_L, which effectively creates an optical oscillator at frequency $\Delta = \omega_c - \omega_L - \delta_{rp}$ (in a reference frame rotating at ω_L; δ_{rp} is the mean shift of the cavity frequency due to radiation pressure). Each of the two oscillators decoheres into its local environment: the optical field at the cavity amplitude decay rate κ and the mechanics at the damping rate γ_m. Entering the desired strong coupling regime requires a coupling rate $g \gtrsim \kappa, \gamma_m$.

The fundamental optomechanical radiation-pressure interaction $H_{int} = -\hbar g_0 n_c X_m$ couples the cavity photon number n_c to the position X_m of the mechanics ($\hbar$ is $h/2\pi$, where h is Planck's constant). On the single-photon level, this interaction provides an intrinsically nonlinear coupling, where the coupling rate $g_0 = \frac{\omega_c}{L}\sqrt{\frac{\hbar}{m\omega_m}}$ (L, cavity length; m, effective mass) describes the effect of a single photon on the optomechanical cavity. In all currently available optomechanical systems, however, g_0 is well below 100 Hz. Because the corresponding cavity decay rates are typically much larger than 10 kHz, the effect is too small to exploit the strong coupling regime on the single-photon level. For our experiment $g_0 = 2\pi \times 2.7$ Hz, which is smaller than both κ ($2\pi \times 215$ kHz) and γ_m ($2\pi \times 140$ Hz). To circumvent this limitation, we use a strong optical driving field ($\lambda = 1{,}064$ nm), which shifts the optomechanical steady state by means of radiation pressure from vacuum to a mean cavity amplitude α (mean cavity photon number $\langle n_c \rangle = \alpha^2$) and from zero displacement to a mean mechanical displacement β. The resulting effective interaction is obtained by standard mean-field expansion, and resembles two harmonic oscillators that are coupled linearly in their optical and mechanical position quadratures $X_c = (a_c + a_c^\dagger)$ and $X_m = (a_m + a_m^\dagger)$, respectively. This strongly driven optomechanical system is then described by equation (1) (see Supplementary Information):

$$H = \frac{\hbar\Delta}{2}\left(X_c^2 + P_c^2\right) + \frac{\hbar\omega_m}{2}\left(X_m^2 + P_m^2\right) - \hbar g X_c X_m \qquad (1)$$

The effective coupling strength $g = g_0\alpha$ is now enhanced by a factor of $\alpha = \sqrt{\langle n_c \rangle}$. Note that this enhancement comes at the cost of losing the nonlinear character of the interaction. Although there exist proposals that do require strong nonlinear coupling at the single-photon level[16], the majority of schemes for quantum optomechanical state manipulation work well within the regime of linear albeit strong coupling. They rely on the fact that linear interactions allow for protocols such as quantum state transfer and readout[23], generation of entanglement[15,17], conditional preparation of states via projective measurements on

[1]Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria. [2]Faculty of Physics, University of Vienna, Strudlhofgasse 4, A-1090 Vienna, Austria. [3]Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Technikerstraße 21a, A-6020 Innsbruck, Austria. [4]Institute for Theoretical Physics, University of Innsbruck, Technikerstrasse 25, A-6020 Innsbruck, Austria.

©2009 Macmillan Publishers Limited. All rights reserved

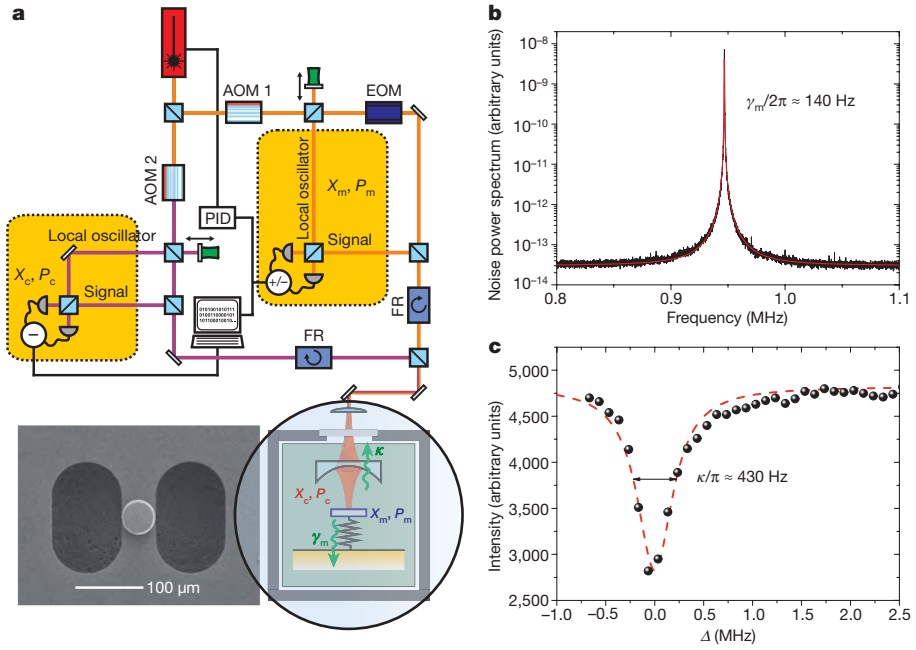


Figure 1 | Experimental set-up and characterization of the uncoupled mechanical and optical oscillator. a, Our micromechanical resonator with a high-reflectivity mirror pad ($R > 0.99991$) that forms the end-face of a 25-mm-long Fabry–Pérot cavity (magnified view circled, bottom right). A strong continuous-wave Nd:YAG laser is used to drive the optomechanical system (purple beam). By splitting off a faint part (15 μW) of the drive laser, the laser frequency is actively locked to the Fabry–Pérot cavity frequency (orange beam). Locking is achieved by phase-modulation (electro-optical modulator, EOM) and by obtaining a Pound-Drever-Hall error signal required for feedback with a proportional–integral–derivative controller (PID). Acousto-optical modulators (AOM) control the relative frequency detuning Δ and thus allow for off-resonant driving of the cavity. Data presented here have been taken by varying the detuning Δ and the power of the drive beam. Both beams are coupled to the Fabry–Pérot cavity via the same spatial mode but orthogonal in polarization. The measured cavity linewidth (full-width at half-maximum, FWHM) $2\kappa \approx 2\pi \times 430$ kHz corresponds to an optical finesse $F \approx 14{,}000$. The fundamental mechanical mode of the microresonator at $\omega_m = 2\pi \times 947$ kHz has a natural linewidth (FWHM) of $\gamma_m \approx 2\pi \times 140$ Hz (mechanical quality factor $Q \approx 6{,}700$) at room temperature. With $\kappa/\omega_m \approx 0.2$, these parameters place us well into the resolved sideband regime $\kappa/\omega_m \ll 1$. The effective mass of 145 ng was obtained by direct fitting of the optomechanical response at low driving powers. After interaction with the optomechanical system, both (drive and lock) beams are separated by a polarizing beamsplitter and Faraday rotators (FR) and are each independently measured by optical homodyning (Supplementary Information). Each homodyne phase can be either scanned or locked to a fixed value by actuating a piezo-driven mirror. **b**, Mechanical noise power spectrum obtained by homodyne detection of the lock beam. Red line, fit to the data assuming an ideal harmonic oscillator in thermal equilibrium. **c**, Intensity of the drive beam that is reflected off the Fabry–Pérot cavity when scanning its detuning Δ, which provides direct access to the cavity transfer function. Dashed red line, Lorentzian fit to the data.

light[18,21], and so on, a fact which is well established in the fields of quantum optics and quantum information. In our experiment, by using external optical pump powers of up to 11 mW, we are able to achieve an increase in coupling by more than five orders of magnitude, sufficient to reach the desired strong coupling regime.

An unambiguous signature of strongly coupled systems is the occurrence of normal mode splitting, a phenomenon known to both classical and quantum physics. In the simplest case, two independent harmonic oscillators coupled via an additional joint spring will behave as a pair of uncoupled oscillators—so-called normal modes—with shifted resonance frequencies compared to the individual resonators. For the particular case of resonators with equal bare frequencies, a sufficiently strong coupling will introduce a spectral splitting of the two normal modes that is of the order of the coupling strength g. Normal mode splitting has been observed in a number of realizations of cQED, where it is also known as Rabi-splitting, with photons coupled either to atoms[24,25,26], to excitons in semiconductor structures[27,28,29] or to Cooper pair box qubits in circuit QED[14]. In case of the strongly driven optomechanical system described by equation (1), the normal modes occur at frequencies $\omega_\pm^2 = \frac{1}{2}(\Delta^2 + \omega_m^2 \pm \sqrt{(\Delta^2 - \omega_m^2)^2 + 4g^2\omega_m\Delta})$ and exhibit a splitting of $\omega_+ - \omega_- \approx g$. In the given simple expression for normal mode frequencies, cavity decay and mechanical damping are neglected. A more careful analysis is carried out in the Supplementary Information, and shows that normal mode splitting occurs only above a threshold $g \gtrsim \kappa$ (refs 19, 20) for our damped optomechanical system. The Hamiltonian can be re-written in terms of the normal modes and one obtains:

$$H = \frac{\hbar\omega_+}{2}\left(X_+^2 + P_+^2\right) + \frac{\hbar\omega_-}{2}\left(X_-^2 + P_-^2\right) \qquad (2)$$

For the resonant case $\Delta = \omega_m$, equation (2) describes two uncoupled oscillators with position and momentum quadratures $X_\pm = \sqrt{\frac{\omega_m \pm g}{2\omega_m}}(X_c \pm X_m)$ and $P_\pm = \sqrt{\frac{\omega_m}{2(\omega_m \pm g)}}(P_c \pm P_m)$. These new dynamical variables cannot be ascribed to either the cavity field or the mechanical resonator, but are true hybrid optomechanical degrees of freedom. The overall system energy spectrum $E_{m,n}$ is therefore given by the sum of the energies of the two normal modes, that is, $E_{m,n} = \hbar(m\omega_+ + n\omega_-)$. The degeneracy of the uncoupled energy levels is lifted, and normal mode splitting of adjacent levels occurs with a separation that is equivalent to the coupling strength g. In the presence of decoherence, the spectral lines are broadened to a width of $(\kappa + \gamma_m)$ and the splitting can therefore only be resolved for $g \gtrsim \kappa, \gamma_m$, that is, for strong coupling.

We observe normal mode splitting via direct spectroscopy of the optical field emitted by the cavity. Emission of a cavity photon can in general be understood as a transition between dressed states of the optomechanical system, that is, between mechanical states that are dressed by the cavity radiation field. The structure of the optomechanical interaction only allows for transitions that lower or raise the total number of normal mode excitations by one (see Supplementary Information). Photons emitted from the cavity therefore

©2009 Macmillan Publishers Limited. All rights reserved

have to lie at sidebands equal to the dressed state frequencies $\omega_{\pm}$ relative to the incoming laser photons of frequency ω_L, that is, they have to be emitted at sideband frequencies $\omega_L \pm \omega_+$ or $\omega_L \pm \omega_-$. Homodyne detection provides us with direct access to the optical sideband spectrum, which is presented in Fig. 2a for the resonant case $\Delta \approx \omega_m$. For small optical pump power, that is, in the regime of weak coupling, the splitting cannot be resolved and one obtains the well-known situation of resolved sideband laser cooling, in which Stokes and anti-Stokes photons are emitted at one specific sideband frequency. The splitting becomes clearly visible at larger pump powers, which is unambiguous evidence for entering the strong coupling regime. Indeed, at a maximum optical driving power of ~11 mW, we obtain a coupling strength $g = 2\pi \times 325$ kHz, which is larger than both $\kappa = 2\pi \times 215$ kHz and $\gamma_m = 2\pi \times 140$ Hz and which corresponds to the magnitude of the level crossing shown in Fig. 2b. As is expected, for detunings Δ off resonance, the normal mode frequencies approach the values of the uncoupled system.

These characteristics of our strongly driven optomechanical system are reminiscent of a strongly driven two-level atom, and indeed a strong and instructive analogy exists. If an atom is pumped by a strong laser field, optical transitions can only occur between dressed atomic states, that is, atomic states 'dressed' by the interaction with the laser field. For strong driving, any Rabi splitting that is induced by strong coupling is effectively of order $G_0\sqrt{\langle n_L \rangle}$ (n_L, mean number of laser photons; G_0, electric dipole coupling) and one therefore obtains an equally spaced level splitting, fully analogous to the coupled optomechanical spectrum. From this point of view, the optomechanical modes can be interpreted in a dressed state approach as excitations of mechanical states that are dressed by the cavity radiation field. The origin of the sideband doublet as observed in the output field of the strongly driven optomechanical cavity corresponds to the resonance fluorescence spectrum of a strongly driven atom, in which strong coupling gives rise to the two side-peaks in the so-called Mollow triplet. It is interesting to note that the analogy even holds for the single-photon regime, in which both systems are close to their quantum ground state. For both cases (that is, the atom–cavity system and the cavity–optomechanical system), a sufficiently strong single-photon interaction g_0 would allow one to obtain the well-known vacuum Rabi splitting as well as state-dependent level spacing, which is due to intrinsic nonlinearities in the coupling.

We should stress that normal mode splitting alone does not establish a proof for coherent dynamics, that is, for quantum interference effects. With the present experimental parameters, such effects are washed out by thermal decoherence and normal mode splitting has a classical explanation in the framework of linear dispersion theory[30]. Still, the demonstration of normal mode splitting is a necessary condition for future quantum experiments.

We finally comment on the prospects for mechanical quantum state manipulation in the regime of strong coupling. One important additional requirement in most proposed schemes is the initialization of the mechanical device close to its quantum ground state. Recent theoretical results show that both ground state laser cooling and strong coupling can be achieved simultaneously, provided that the conditions $\frac{k_B T}{\hbar Q} \ll \kappa \ll \omega_m$ are fulfilled[20,22]. Thus, in addition to operating in the resolved sideband regime, a thermal decoherence rate that is small compared to the cavity decay rate is required. Cryogenic experiments have demonstrated thermal decoherence rates as low as 20 kHz for nanomechanical resonators for a 20 mK environment temperature[9]. For our experiment, temperatures below 300 mK would be sufficient to combine strong coupling with ground state cooling.

We have demonstrated strong coupling of a micromechanical resonator to an optical cavity field. This regime is a necessary precondition to obtaining quantum control of mechanical systems.

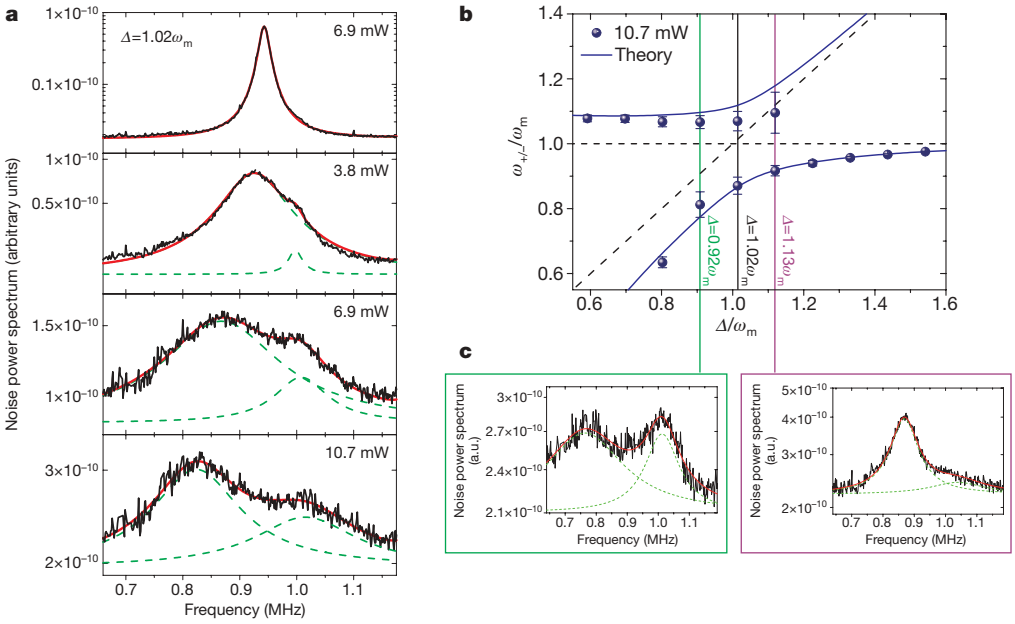


Figure 2 | Optomechanical normal mode splitting and avoided crossing in the normal-mode frequency spectrum. **a**, Emission spectra of the driven optomechanical cavity, obtained from sideband homodyne detection on the strong driving field after its interaction with the optomechanical system (see Supplementary Information). The power levels from top to bottom (0.6, 3.8, 6.9, 10.7 mW) correspond to an increasing coupling strength of $g = 78$, 192, 260 and 325 kHz ($g = 0.4$, 0.9, 1.2, 1.5 κ). All measurements are performed close to resonance ($\Delta = 1.02\,\omega_m$). For strong driving powers a splitting of the cavity emission occurs, corresponding to the normal mode frequencies of true hybrid optomechanical degrees of freedom. This normal mode splitting is an unambiguous signature of the strong coupling regime. All plots are shown on a logarithmic scale. Green dashed lines are fits to the data assuming two independent Lorentzian curves, red solid lines are the sum signal of these two fits. **b**, Normal mode frequencies obtained from the fits to the spectra as a function of detuning Δ. For far off-resonant driving, the normal modes approach the limiting case of two uncoupled systems. Dashed lines indicate the frequencies of the uncoupled optical (diagonal) and mechanical (horizontal) resonator, respectively. At resonance, normal mode splitting prevents a frequency degeneracy, which results in the shown avoided level crossing. Error bars, s.d. Solid lines are simulations (see Supplementary Information). For larger detuning values, the second normal mode peak could no longer be fitted owing to a nearby torsional mechanical mode. **c**, Normal mode spectra measured off resonance.

©2009 Macmillan Publishers Limited. All rights reserved

Together with the availability of high-quality mechanical resonators operated at low temperatures, which minimizes thermal decoherence of the mechanics, strong optomechanical coupling provides the basis for full photonic quantum control of massive mechanical resonators. We suggest that future developments will eventually also allow strong coupling to be achieved in the nonlinear regime, that is, at the single-photon level[11,16], to exploit optomechanical vacuum Rabi splitting.

Received 4 February; accepted 26 May 2009.

1. Naik, A. *et al.* Cooling a nanomechanical resonator with quantum back-action. *Nature* **443,** 193–196 (2006).
2. Cleland, A. N., Aldridge, J. S., Driscoll, D. C. & Gossard, A. C. Nanomechanical displacement sensing using a quantum point contact. *Appl. Phys. Lett.* **81,** 1699–1701 (2002).
3. Rugar, D., Budakian, R., Mamin, H. J. & Chui, B. W. Single spin detection by magnetic resonance force microscopy. *Nature* **430,** 329–332 (2004).
4. Rabl, P. *et al.* Strong magnetic coupling between an electronic spin qubit and a mechanical resonator. *Phys. Rev. B* **79,** 041302(R) (2009).
5. Kippenberg, T. J., Rokhsari, H., Carmon, T., Scherer, A. & Vahala, K. J. Analysis of radiation-pressure induced mechanical oscillation of an optical microcavity. *Phys. Rev. Lett.* **95,** 033901 (2005).
6. Gigan, S. *et al.* Self-cooling of a micromirror by radiation pressure. *Nature* **444,** 67–71 (2006).
7. Arcizet, O., Cohadon, P.-F., Briant, T., Pinard, M. & Heidmann, A. Radiation-pressure cooling and micromechanical instability of a micromirror. *Nature* **444,** 71–75 (2006).
8. Thompson, J. D. *et al.* Strong dispersive coupling of a high-finesse cavity to a micromechanical membrane. *Nature* **452,** 72–75 (2008).
9. Regal, C. A., Teufel, J. D. & Lehnert, K. W. Measuring nanomechanical motion with a microwave cavity interferometer. *Nature Phys.* **4,** 555–560 (2008).
10. Eichenfield, M., Michael, C. P., Perahia, R. & Painter, O. Actuation of micro-optomechanical systems via cavity-enhanced optical dipole forces. *Nature Photon.* **1,** 416–422 (2007).
11. Li, M. *et al.* Harnessing optical forces in integrated photonic circuits. *Nature* **456,** 480–484 (2008).
12. Walther, H., Varcoe, B. T. H., Englert, B.-G. & Becker, T. Cavity quantum electrodynamics. *Rep. Prog. Phys.* **69,** 1325–1382 (2006).
13. Khitrova, G., Gibbs, H. M., Kira, M., Koch, S. W. & Scherer, A. Vacuum Rabi splitting in semiconductors. *Nature Phys.* **2,** 81–90 (2006).
14. Wallraff, A. *et al.* Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431,** 162–167 (2004).
15. Bose, S., Jacobs, K. & Knight, P. L. Preparation of nonclassical states in cavities with a moving mirror. *Phys. Rev. A* **56,** 4175–4186 (1997).
16. Marshall, W., Simon, C., Penrose, R. & Bouwmeester, D. Towards quantum superpositions of a mirror. *Phys. Rev. Lett.* **91,** 130401 (2003).
17. Vitali, D. *et al.* Optomechanical entanglement between a movable mirror and a cavity field. *Phys. Rev. Lett.* **98,** 030405 (2007).
18. Clerk, A. A., Marquardt, F. & Jacobs, K. Back-action evasion and squeezing of a mechanical resonator using a cavity detector. *N. J. Phys.* **10,** 095010 (2008).
19. Marquardt, F., Chen, J. P., Clerk, A. A. & Girvin, S. M. Quantum theory of cavity-assisted sideband cooling of mechanical motion. *Phys. Rev. Lett.* **99,** 093902 (2007).
20. Dobrindt, J. M., Wilson-Rae, I. & Kippenberg, T. J. Parametric normal-mode splitting in cavity optomechanics. *Phys. Rev. Lett.* **101,** 263602 (2008).
21. Hammerer, K., Aspelmeyer, M., Polzik, E. & Zoller, P. Establishing Einstein-Podolsky-Rosen channels between nanomechanics and atomic ensembles. *Phys. Rev. Lett.* **102,** 020501 (2009).
22. Wilson Rae, I., Nooshi, N., Dobrindt, J., Kippenberg, T. J. & Zwerger, W. Cavity-assisted backaction cooling of mechanical resonators. *N. J. Phys.* **10,** 095007 (2008).
23. Zhang, J., Peng, K. & Braunstein, S. L. Quantum-state transfer from light to macroscopic oscillators. *Phys. Rev. A* **68,** 013808 (2003).
24. Thompson, R. J., Rempe, G. & Kimble, H. J. Observation of normal-mode splitting for an atom in an optical cavity. *Phys. Rev. Lett.* **68,** 1132–1135 (1992).
25. Colombe, Y. *et al.* Strong atom–field coupling for Bose–Einstein condensates in an optical cavity on a chip. *Nature* **450,** 272–276 (2007).
26. Aoki, T. *et al.* Observation of strong coupling between one atom and a monolithic microresonator. *Nature* **443,** 671–674 (2006).
27. Reithmaier, J. P. *et al.* Strong coupling in a single quantum dot–semiconductor microcavity system. *Nature* **432,** 197–200 (2004).
28. Weisbuch, C., Nishioka, M., Ishikawa, A. & Arakawa, Y. Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Phys. Rev. Lett.* **69,** 3314–3317 (1992).
29. Yoshie, T. *et al.* Vacuum Rabi splitting with a single quantum dot in a photonic crystal nanocavity. *Nature* **432,** 200–203 (2004).
30. Zhu, Y. *et al.* Vacuum Rabi splitting as a feature of linear-dispersion theory: analysis and experimental observations. *Phys. Rev. Lett.* **64,** 2499–2502 (1990).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to T. Corbitt, C. Genes, S. Goßler, P. K. Lam, G. Milburn, P. Rabl and P. Zoller for discussions. We also thank M. Metzler, R. Ilic and M. Skvarla (CNF), and K. C. Schwab and J. Hertzberg, for microfabrication support, and R. Blach for technical support. We acknowledge financial support from the Austrian Science Fund FWF, the European Commission and the Foundational Questions Institute. S.G. is a recipient of a DOC fellowship of the Austrian Academy of Sciences; S.G. and M.R.V. are members of the FWF doctoral programme Complex Quantum Systems (CoQuS).

Author Contributions All authors have made a significant contribution to the concept, design, execution or interpretation of the presented work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.A. (markus.aspelmeyer@quantum.at).

©2009 **Macmillan Publishers Limited. All rights reserved**

Vol 460|6 August 2009|doi:10.1038/nature08213

LETTERS

The late Precambrian greening of the Earth

L. Paul Knauth[1] & Martin J. Kennedy[2]

Many aspects of the carbon cycle can be assessed from temporal changes in the $^{13}C/^{12}C$ ratio of oceanic bicarbonate. $^{13}C/^{12}C$ can temporarily rise when large amounts of ^{13}C-depleted photosynthetic organic matter are buried at enhanced rates[1], and can decrease if phytomass is rapidly oxidized[2] or if low ^{13}C is rapidly released from methane clathrates[3]. Assuming that variations of the marine $^{13}C/^{12}C$ ratio are directly recorded in carbonate rocks, thousands of carbon isotope analyses of late Precambrian examples have been published to correlate these otherwise undatable strata and to document perturbations to the carbon cycle just before the great expansion of metazoan life. Low $^{13}C/^{12}C$ in some Neoproterozoic carbonates is considered evidence of carbon cycle perturbations unique to the Precambrian. These include complete oxidation of all organic matter in the ocean[2] and complete productivity collapse such that low-$^{13}C/^{12}C$ hydrothermal CO_2 becomes the main input of carbon[4]. Here we compile all published oxygen and carbon isotope data for Neoproterozoic marine carbonates, and consider them in terms of processes known to alter the isotopic composition during transformation of the initial precipitate into limestone/dolostone. We show that the combined oxygen and carbon isotope systematics are identical to those of well-understood Phanerozoic examples that lithified in coastal pore fluids, receiving a large groundwater influx of photosynthetic carbon from terrestrial phytomass. Rather than being perturbations to the carbon cycle, widely reported decreases in $^{13}C/^{12}C$ in Neoproterozoic carbonates are more easily interpreted in the same way as is done for Phanerozoic examples. This influx of terrestrial carbon is not apparent in carbonates older than ~850 Myr, so we infer an explosion of photosynthesizing communities on late Precambrian land surfaces. As a result, biotically enhanced weathering generated carbon-bearing soils on a large scale and their detrital sedimentation sequestered carbon[5]. This facilitated a rise in O_2 necessary for the expansion of multicellular life.

Carbonate is initially precipitated in the ocean as metastable calcite and/or aragonite, and is subsequently transformed into stable interlocking crystals of low-Mg calcite and/or dolomite. In this 'lithification' process, the originally precipitated phases undergo dissolution, reprecipitation, and isotopic re-equilibration with ambient pore fluids[6]. With the exception of rare, relatively well-preserved Phanerozoic shells, this happens to all carbonates. Precambrian carbonates are limestone or dolostone rocks; there are no preserved initial Precambrian marine precipitates.

On the basis of studies of Cenozoic deposits, lithification is most rapid in sediments in coastal areas where meteoric ground waters mix with marine pore fluids[6]. The process yields a roughly co-variant relationship between $\delta^{13}C$ and $\delta^{18}O$ in which $\delta^{13}C$ decreases as $\delta^{18}O$ decreases[7–9] (Fig. 1a). The trend documents phases that formed in mixed meteoric/marine pore fluids. Meteoric waters are depleted in ^{18}O relative to ocean water, so $\delta^{18}O$ decreases with increasing proportions of meteoric water. Photosynthesizing communities living in coastal recharge areas preferentially fix ^{12}C, such that the biomass is typically over 20‰ depleted in ^{13}C relative to the marine inorganic C reservoir. This ^{13}C-depleted C is incorporated into the meteoric water

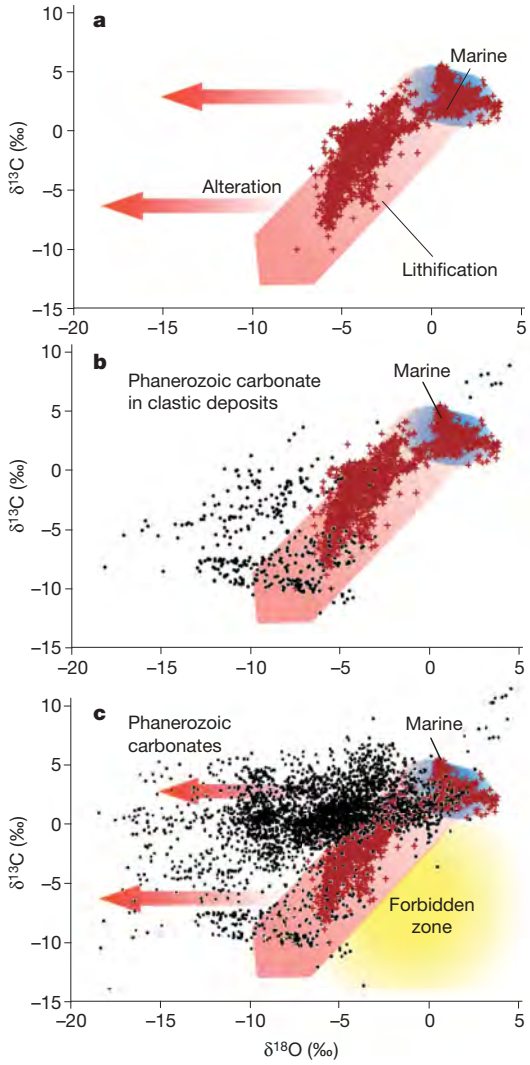


Figure 1 | Stable isotopes in Phanerozoic carbonates. a, Cenozoic limestones/dolostones form by way of solution/reprecipitation of primary marine precipitates. Blue area is dominantly marine pore fluids, red area contains significant meteoric water component. Both areas define the 'lithification' zone. Later deep burial and/or metamorphic alteration yield data to the left of the lithification domain, as shown by the arrows. **b**, Carbonate cements, lenses and beds in dominantly clastic successions. The lithification domain from **a** is shown in this and all subsequent figure panels for reference. **c**, Phanerozoic carbonates plot in or to the left of the lithification zone and avoid the forbidden zone. Data for this and all other figures are available in Supplementary Information.

[1]School of Earth and Space Exploration, Arizona State University, Tempe, Arizona 85287-1404, USA. [2]Department of Earth Science, University of California, Riverside, Riverside, California 92557, USA.

©2009 Macmillan Publishers Limited. All rights reserved

pore fluids through decomposition of the constantly renewing biomass and becomes incorporated into limestone/dolostone during lithification. A co-variation between $\delta^{18}O$ and $\delta^{13}C$ arises because low $\delta^{18}O$ signifies more meteoric water and thus a greater introduction of photosynthetically depleted ^{13}C from the subaerial recharge area. Pervasive intrusion of meteoric waters during sea-level fall is common on carbonate platforms. Pleistocene examples yield negative $\delta^{13}C$ and $\delta^{18}O$ over stratigraphic thicknesses >100 m (ref. 7). $\delta^{13}C$ systematically decreases upward in the stratigraphic succession, like patterns widely reported for Neoproterozoic strata. The stratigraphic trend develops during lithification and is unrelated to any change in the global oceanic $\delta^{13}C$.

Carbonate rocks are susceptible to later recrystallization during deep burial, and can isotopically re-equilibrate with higher-temperature fluids depending upon the water/rock (W/R) ratios and the isotopic composition of the fluids. The W/R ratio of metamorphic fluids is typically high with respect to O and low with respect to C, so carbonate $\delta^{18}O$ is driven to lower values through higher-temperature equilibration while $\delta^{13}C$ undergoes little change[10] (Fig. 1a). Ancient carbonates that form in coastal pore fluids and undergo later alteration therefore display values that plot in, or to the left of, the lithification domain (Fig. 1a). Whereas $\delta^{13}C$ values as low as −10‰ can occur, most are greater than −5‰. Low $\delta^{18}O$ of an ancient rock may have been set during the initial lithification event or during a much later metamorphic event. For example, a rock sample in the lithification domain (Fig. 1a) with $\delta^{18}O = -4$‰, $\delta^{13}C = -3$‰ does not represent the value of the initially precipitated marine sediment but does represent the initial, 'unaltered' value of the limestone or dolostone that formed during lithification. In Cenozoic rocks, its position in the crossplot would indicate that meteoric waters rich in coastal photosynthetic C were a component of the pore fluids during lithification. Neoproterozoic carbonate rocks should first be evaluated in this manner, which is well understood for Cenozoic limestone/dolostone formation.

Many published Neoproterozoic C isotope data are for dominantly clastic sections with relatively thin carbonate beds, cements and accumulations. Phanerozoic examples of carbonate in dominantly clastic sequences also display wide variations in $\delta^{18}O$ and $\delta^{13}C$ (Fig. 1b). In some successions, the isotopic signal is set during lithification in isotopically evolving pore fluids, as discussed above.

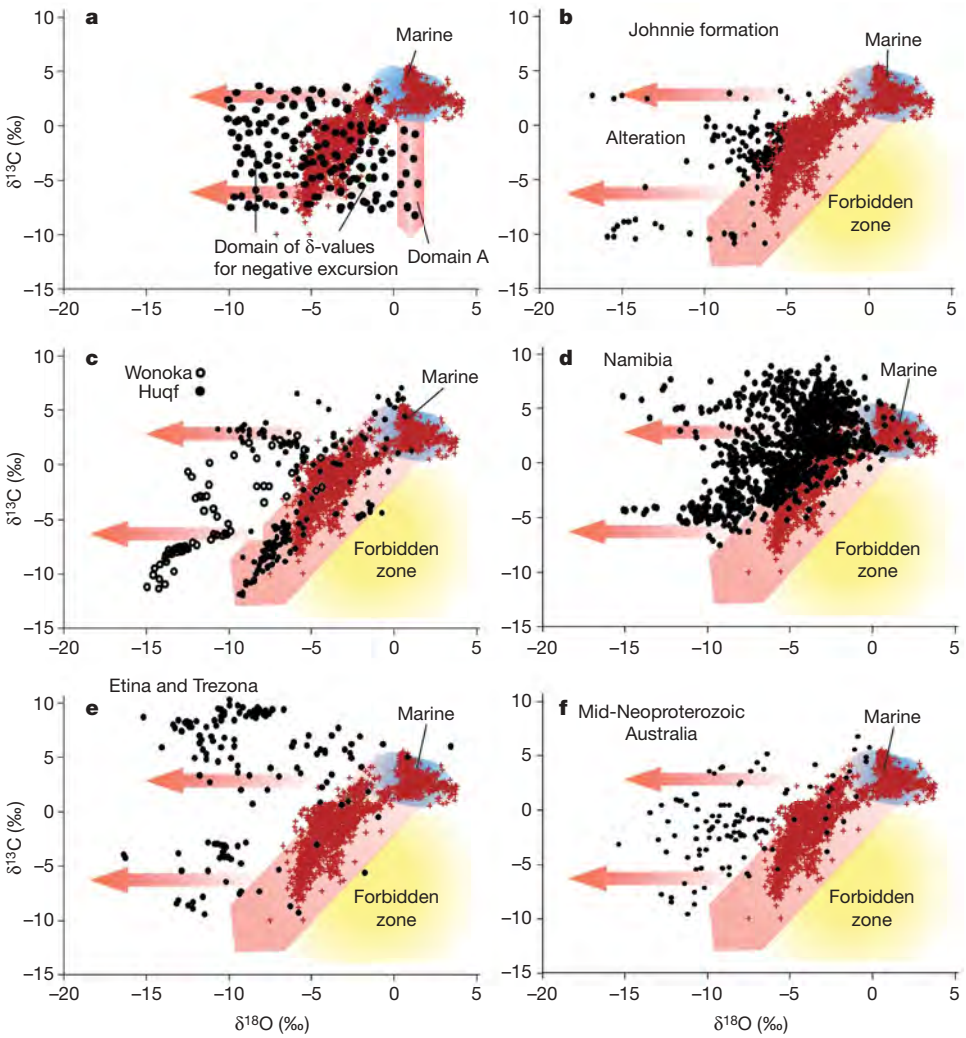


Figure 2 | Neoproterozoic carbonates. **a**, Data pattern expected for marine carbonates during a negative excursion of the global marine inorganic C reservoir. $\delta^{18}O$ of sea water is unaffected, so initial precipitates have marine $\delta^{18}O$ and progressively lower $\delta^{13}C$ depending upon the magnitude of the negative excursion (domain A). Subsequent alteration shifts the values to lower $\delta^{18}O$. No forbidden zone occurs. **b**–**f**, Data for claimed negative C excursions all plot in or to the left of the forbidden zone. **b**, Johnnie Formation, USA; **c**, Wonoka Formation (Australia) and Huqf Group (Oman); **d**, Namibia (Africa); **e**, Etina and Trezona Formations (Australia); and **f**, Mid-Neoproterozoic (Australia).

©2009 Macmillan Publishers Limited. All rights reserved

For example, low $\delta^{18}O$ in carbonate-cemented shelf sands and silts results from meteoric waters in phreatic lenses, and low $\delta^{13}C$ is contributed from the terrestrial phytomass[11]. More commonly, circulating deep basinal fluids produce late carbonate cements and lenses during burial. These have low $\delta^{18}O$ from equilibration with pore fluids at elevated temperatures, and $\delta^{13}C$ typically less than −5‰ from thermal decarboxylation of buried organic matter encountered along the fluid flow path[12,13]. In all cases, the data plot near or to the left of the lithification reference domain (Fig. 1b).

In clastic-dominated sequences, the pore fluid W/R ratio for C is much larger than in limestone/dolostone sections. Any initial carbonate precipitates are therefore more susceptible to alteration of the C isotope ratio. Modern phreatic lenses over 100 m thick occur today in clastic sequences out to 100 km offshore[14]. Submarine groundwater flux rates through these lenses even exceed river runoff[15]. Such lenses migrate as a band across the shelf that tracks sea-level variation and exposes the vast majority of clastic shelf sediment to meteoric waters at some point during burial[11,16]. Shelf sediments dominate past successions preserved in the stratigraphic record, so thick clastic sequences with low-^{18}O, low-^{13}C carbonate are expected.

The pattern of lithification followed by possible later metamorphism results in a forbidden zone, which contains little data, located to the right of the lithification zone on the isotopic crossplot. Thousands of published analyses of Phanerozoic carbonates clearly avoid the forbidden zone (Fig. 1c). The majority of these data are for carbonate-dominated systems and thus have $\delta^{13}C > -5‰$.

Thousands of $\delta^{13}C$ analyses have now been published for Neoproterozoic carbonates to possibly correlate strata and explore for perturbations to the global C cycle. As observed in the Pleistocene[7], thick intervals of negative $\delta^{13}C$ values occur in the Neoproterozoic but are considered to record profound perturbations to the global marine carbon cycle unique to the Neoproterozoic rather than being considered in terms of normal mixed meteoric/marine lithification or decarboxylation. In some cases, stratigraphic relations require these excursions to be sustained for millions of years at values even below the mantle average of −5‰, the minimum value for a steady state C isotope mass balance in the Phanerozoic ocean[17].

The relationship between $\delta^{13}C$ and $\delta^{18}O$ provides a useful means of distinguishing meteoric and burial lithification from marine excursions. Putative changes in the $\delta^{13}C$ of the ocean relate only to the well-mixed dissolved inorganic C reservoir, and show no relationship to $\delta^{18}O$ because the $\delta^{18}O$ of sea water is affected by totally different processes. A perturbation to the global C cycle that lowers the $\delta^{13}C$ of the surface ocean's dissolved inorganic C reservoir by claimed values of 5–12‰ (for example, ref. 2) therefore yields a negative excursion with values that would plot in the forbidden zone. Indeed, such excursions would define a vertical band for the precipitated carbonate as the excursion progressed down from the normal marine carbonate values (Fig. 2a). If the $\delta^{13}C$ of the precipitate is inviolate as currently assumed, later alteration drives $\delta^{18}O$ to lower values and a 'box-like' data array results, bounded to the right by the vertical excursion trajectory (Fig. 2a, domain A). This array is therefore expected, in contrast with that observed for the Phanerozoic.

Five classic examples of successions claiming to represent marine negative C isotope excursions do not show the expected crossplot for such excursions (Fig. 2b–f). Instead, they are bounded by the Cenozoic lithification domain to the right, and stream off to lower $\delta^{18}O$ values that are readily attributable to later metamorphic alteration. In all cases, the lower $\delta^{13}C$ values are associated with lower $\delta^{18}O$ values.

These data are incompatible with the excursion scenario unless (1) carbonates deposited during a negative $\delta^{13}C$ excursion are more susceptible to later $\delta^{18}O$ metamorphic alteration in proportion to their negative $\delta^{13}C$ value, or (2) $\delta^{18}O$ of sea water decreases as $\delta^{13}C$ of sea water decreases to produce an array that fortuitously resembles the Cenozoic lithification domain (Fig. 1a). There is no known geological mechanism compatible with the first scenario, and the second would require synchronous changes of >8‰ in the $\delta^{18}O$ of sea water during the time of the putative C isotope excursions. No mechanisms that synchronously change oceanic $\delta^{18}O$ are known or have been proposed. The alternative explanation—that the Neoproterozoic carbonates underwent the same processes of lithification and metamorphism observed in Phanerozoic carbonates—is a simpler explanation for these data, and does not require extraordinary and anomalous behaviour of the C cycle specific to the Neoproterozoic.

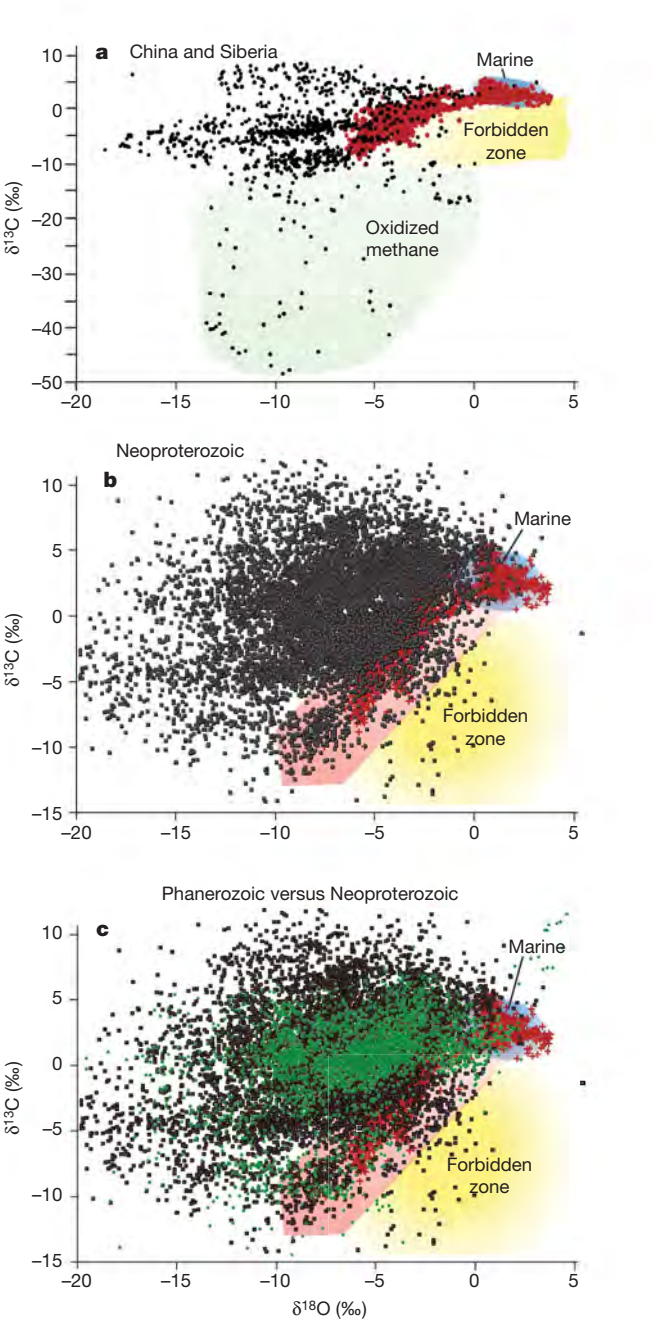


Figure 3 | All Neoproterozoic and Phanerozoic carbonates. a, Extreme ^{13}C depletions associated with oxidation of methane in samples from China and Siberia. Note expanded vertical scale. **b**, All published Neoproterozoic data. Nearly all plot in or to the left of the lithification reference domain. Most of the few values plotting in the forbidden zone probably represent unusual examples involving oxidation of methane, as documented for the China and Siberia forbidden zone samples. **c**, Comparison of Phanerozoic carbonates (green points) with Neoproterozoic carbonates. The data are isotopically indistinguishable.

©2009 Macmillan Publishers Limited. All rights reserved

Some Neoproterozoic carbonates with $\delta^{13}C$ as low as −40‰ do plot in the forbidden zone (Fig. 3a), but most of these have been attributed to local oxidation of methane[3,18]. All unfiltered, published Neoproterozoic data (over 8,000 samples; references in Supplementary Information) yield very few samples in the forbidden zone (Fig. 3b). Several of those that do are thin horizons in otherwise low-^{13}C intervals, and several are for claimed glaciomarine concretions[19]. The Neoproterozoic data are essentially indistinguishable from the Phanerozoic data (Fig. 3c). Similar to their Phanerozoic counterparts, carbonate sequences have $\delta^{13}C > -5‰$, whereas clastic sequences yield $\delta^{13}C$ as low as −10‰. Rather than indicating C isotope excursions, the Neoproterozoic data seem to represent normal isotope variations that occur during lithification and metamorphism. Apparent correlations of low $\delta^{13}C$ between widely separated sections may relate simply to eustatic sea-level drops where phreatic lenses intrude simultaneously during lithification.

Consequences of this interpretation include: (1) $\delta^{18}O$ values of Neoproterozoic sea water were similar to that of the Cenozoic, consistent with the Muehlenbachs model[20]; (2) a terrestrial phytomass existed that was large enough to inject enough photosynthetic C into coastal ground waters to produce identical C isotope depletions in marine carbonate to those observed in the Phanerozoic. The nature of the photosynthesizers is obscured by the absence of preserved terrestrial surfaces, although molecular and other evidence suggest the expansion of a primitive land biota starting at least by 1 Gyr ago[5,21–23]. It was probably composed of protists, mosses, fungi and liverworts with evidence of lichen by 600 Myr ago[5,24]. (3) Marine $\delta^{13}C$ may have varied in the Neoproterozoic, but this could only be deduced following careful consideration of large isotopic changes that occur during lithification and later alteration. The most positive $\delta^{13}C$ in a co-variant data array is the best minimum estimate of the marine $\delta^{13}C$ of the local marine depositional environment. This approach was used in one early study[25], and similar analysis is now warranted for all other Neoproterozoic sections.

If the first expansion of photosynthesizing communities on the land surface occurred in the Neoproterozoic, then carbonates deposited before 850 Myr ago should not show significant negative $\delta^{13}C$ values. Indeed, they cluster around $\delta^{13}C = 0 \pm 2‰$ (Fig. 4). Any earlier coastal phytomass was apparently not significant enough to produce strongly negative $\delta^{13}C$ in carbonates during the stabilization process.

The contrasting isotope data between 850 Myr ago and the Neoproterozoic suggest that the terrestrial expansion of photosynthesizing communities preceded the significant climate perturbations of the late Precambrian glaciations, and was followed by a rise of O_2 (ref. 26) and a secular change in terrestrial sediment composition[5]. The onset of significant biotically enhanced terrestrial weathering would have increased the flux of lithophile nutrient elements and clay minerals to continental margins. This would have increased production and burial preservation of organic C towards modern values[5,27,28] and consequently facilitated the stepwise rise in atmospheric O_2 necessary to support multicellularity. The terrestrial expansion of an extensive, simple land biota indicated by the isotope data may thus have been a critical step in the transition from the Precambrian to the Phanerozoic world.

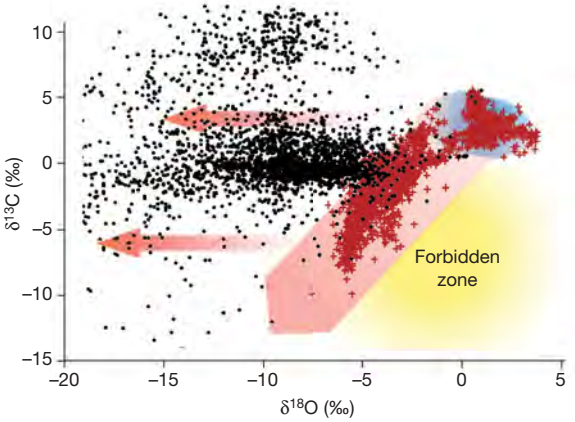


Figure 4 | Pre-850 Myr ago marine calcite and dolomite. These data do not display the ^{13}C depletions associated with lithification as observed for younger carbonates. Low ^{13}C was not being introduced during lithification, indicating minimal photosynthesizing communities on land in coastal areas. Being much older, these have undergone significantly more metamorphism to lower $\delta^{18}O$ values. Most of the very low $\delta^{13}C$ values are for vein-rich samples or high-grade metamorphic rocks.

METHODS SUMMARY

All data plotted in the figures were extracted entirely from the published literature and online databases. All known Neoproterozoic data published as of 2008 were compiled without filtering or exclusion. The accompanying Excel file (Supplementary Information) lists all data and references for convenient reconstruction of the figures.

Received 20 June 2008; accepted 18 June 2009.
Published online 8 July 2009.

1. Scholle, P. A. & Arthur, M. A. Carbon isotope fluctuations in Cretaceous pelagic limestones: potential stratigraphic and petroleum exploration tool. *Am. Assoc. Petrol. Geol. Bull.* **64,** 67–87 (1980).
2. Fike, D. A., Grotzinger, J. P., Pratt, L. M. & Summons, R. E. Oxidation of the Ediacaran Ocean. *Nature* **444,** 744–747 (2006).
3. Jiang, G. Q., Kennedy, M. J. & Christie-Blick, N. Stable isotopic evidence for methane seeps in Neoproterozoic postglacial cap carbonates. *Nature* **426,** 822–826 (2003).
4. Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A Neoproterozoic snowball earth. *Science* **281,** 1342–1346 (1998).
5. Kennedy, M., Droser, M., Mayer, L. M., Pevear, D. & Mrofka, D. Late Precambrian oxygenation; inception of the clay mineral factory. *Science* **311,** 1446–1449 (2006).
6. Land, L. S. Limestone diagenesis — some geochemical considerations. *US Geol. Surv. Bull.* **1578,** 129–137 (1986).
7. Melim, L. A., Swart, P. K. & Maliva, R. G. in *Subsurface Geology of a Prograding Carbonate Platform Margin, Great Bahama Bank: Results of the Bahamas Drilling Project* Vol. 70 (ed. Ginsburg, R.N.) 137–161 (SEPM, 2001).
8. Quinn, T. M. Meteoric diagenesis of Plio-Pleistocene limestones at Enewetak Atoll. *J. Sedim. Petrol.* **61,** 681–703 (1990).
9. Gross, M. G. & Tracey, J. I. Oxygen and carbon isotopic composition of limestones and dolomites, Bikini and Eniwetok Atolls. *Science* **151,** 1082–1084 (1966).
10. Banner, J. L. & Hanson, G. N. Calculation of simultaneous isotopic and trace-element variations during water-rock interaction with applications to carbonate diagenesis. *Geochim. Cosmochim. Acta* **54,** 3123–3137 (1990).
11. Taylor, K. G., Gawthorpe, R. L., Curtis, C. D., Marshall, J. D. & Awwiller, D. N. Carbonate cementation in a sequence-stratigraphic framework: Upper Cretaceous sandstones, Book Cliffs, Utah-Colorado. *J. Sedim. Res.* **70,** 360–372 (2000).
12. Hendry, J. P., Wilkinson, M., Fallick, A. E. & Haszeldine, R. S. Ankerite cementation in deeply buried Jurassic sandstone reservoirs of the central North Sea. *J. Sedim. Res.* **70,** 227–239 (2000).
13. Fayek, M. *et al.* In situ stable isotopic evidence for protracted and complex carbonate cementation in a petroleum reservoir, North Coles Levee, San Joaquin Basin, California, USA. *J. Sedim. Res.* **71,** 444–458 (2001).
14. Hathaway, J. C. *et al.* United States Geological Survey core drilling on the Atlantic Shelf. *Science* **206,** 515–527 (1979).
15. Moore, W. S. S. a. r. m. i. e. n. t. o. J. L. & Key, R. M. Submarine groundwater discharge revealed by ^{228}Ra distribution in the upper Atlantic Ocean. *Nature Geosci.* **1,** 309–311 (2008).
16. Brooks, S. M. & Whitaker, F. F. Geochemical and physical controls on vadose zone hydrology of Holocene carbonate sands, grand Bahama Island. *Earth Surf. Process. Landforms* **22,** 45–58 (1997).
17. Bristow, T. F. & Kennedy, M. J. Carbon isotope excursions and the oxidant budget of the Ediacaran atmosphere and ocean. *Geology* **36,** 863–866 (2008).
18. Pokrovskii, B. G., Melezhik, V. A. & Bujakaite, M. I. Carbon, oxygen, strontium, and sulfur isotopic compositions in late Precambrian rocks of the Patom Complex, central Siberia: Communication 2. Nature of carbonates with ultralow and ultrahigh $\delta^{13}C$ values. *Lithol. Miner. Res.* **41,** 576–587 (2006).
19. Fairchild, I. J. & Spiro, B. Carbonate minerals in glacial sediments: geochemical clues to palaeoenvironment. *Geol. Soc. Lond. Spec. Publ.* **53,** 201–216 (1990).
20. Muehlenbachs, K. The oxygen isotopic composition of the oceans, sediments and the seafloor. *Chem. Geol.* **145,** 263–273 (1998).
21. Horodyski, R. J. & Knauth, L. P. Life on land in the Precambrian. *Science* **263,** 494–498 (1994).
22. Heckman, D. S. *et al.* Molecular evidence for the early colonization of land by fungi and plants. *Science* **293,** 1129–1133 (2001).

©2009 Macmillan Publishers Limited. All rights reserved

23. Prave, A. R. Life on land in the Proterozoic: evidence from the Torridonian rocks of northwest Scotland. *Geology* **30,** 811–814 (2002).
24. Yuan, X. L., Xiao, S. H. & Taylor, T. N. Lichen-like symbiosis 600 million years ago. *Science* **308,** 1017–1020 (2005).
25. Kaufman, A. J., Knoll, A. H. & Awramik, S. M. Biostratigraphic and chemostratigraphic correlation of Neoproterozoic sedimentary successions-Upper Tindir Group, Northwestern Canada as a test case. *Geology* **20,** 181–185 (1992).
26. Canfield, D. E. The early history of atmospheric oxygen: homage to Robert A. Garrels. *Annu. Rev. Earth Planet. Sci.* **33,** 1–36 (2005).
27. Lenton, T. M. & Watson, A. J. Biotic enhancement of weathering, atmospheric oxygen and carbon dioxide in the Neoproterozoic. *Geophys. Res. Lett.* **31,** L05202, doi:10.1029/2003GL018802 (2004).
28. Schwartzman, D. *Life, Temperature, and the Earth* (Columbia Univ. Press, 1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. McFadden for help compiling the data. L.P.K. was funded by NASA Exobiology grants NG04GJ47G and NNX08AT72G. M.J.K. was funded by NASA Exobiology NNG04GJ42G and NSF EAR 0345207.

Author Contributions Both authors shared equally in interpretations and implications of the data. L.P.K. compiled the data, wrote the initial draft, and managed revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.P.K. (Knauth@asu.edu).

©2009 **Macmillan Publishers Limited. All rights reserved**

LETTERS

Fluid and deformation regime of an advancing subduction system at Marlborough, New Zealand

Philip E. Wannamaker[1], T. Grant Caldwell[2], George R. Jiracek[3], Virginie Maris[4], Graham J. Hill[2], Yasuo Ogawa[5], Hugh M. Bibby[2], Stewart L. Bennie[2] & Wiebke Heise[2]

Newly forming subduction zones on Earth can provide insights into the evolution of major fault zone geometries from shallow levels to deep in the lithosphere and into the role of fluids in element transport and in promoting rock failure by several modes[1,2]. The transpressional subduction regime of New Zealand, which is advancing laterally to the southwest below the Marlborough strike–slip fault system of the northern South Island[3,4], is an ideal setting in which to investigate these processes. Here we acquired a dense, high-quality transect of magnetotelluric soundings across the system, yielding an electrical resistivity cross-section to depths beyond 100 km. Our data imply three distinct processes connecting fluid generation along the upper mantle plate interface to rock deformation in the crust as the subduction zone develops. Massive fluid release just inland of the trench induces fault-fracture meshes through the crust above that undoubtedly weaken it as regional shear initiates. Narrow strike–slip faults in the shallow brittle regime of interior Marlborough diffuse in width upon entering the deeper ductile domain aided by fluids and do not project as narrow deformation zones. Deep subduction-generated fluids rise from 100 km or more and invade upper crustal seismogenic zones that have exhibited historic great earthquakes on high-angle thrusts that are poorly oriented for failure under dry conditions. The fluid-deformation connections described in our work emphasize the need to include metamorphic and fluid transport processes in geodynamic models.

Modern New Zealand straddles the active Australian–Pacific plate boundary and its present physiography largely reflects ongoing transpressional deformation since the Late Cenozoic era[3,4]. In its North Island, the dextral differential motion is accommodated by oblique subduction of the Pacific plate into the Hikurangi trench with relatively weak plate coupling (Fig. 1). Passing southwestward to the Marlborough district of the northern South Island, the plate-normal component is still taken up mainly by subduction. However, the degree of coupling to upper-plate lithology is greater and is expressed as four major strike–slip faults (Alpine/Wairau, Awatere, Clarence and Hope), plus an incipient fifth fault (Porters Pass). The ages of initiation of the strike–slip faults become progressively younger to the southeast, a reflection of the advance of the subduction system southwestward[3]. Seismicity in the Marlborough region is high, but is

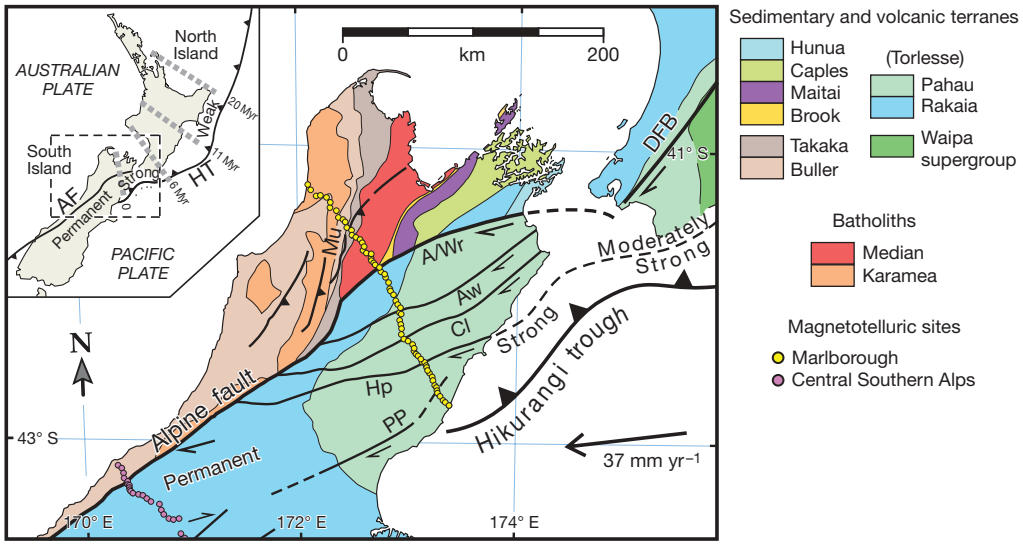


Figure 1 | Geological terrane map of central New Zealand[8]. The Marlborough magnetotelluric transect is shown (yellow circles) and so is the northwest portion of the central Southern Alps magnetotelluric transect (pink circles). The inset shows southwestward migration of the edge of the Hikurangi subduction zone over time[4]. The transition in degree of coupling between subducted Pacific plate and the Australian plate to the northwest is denoted weak, strong and permanent, moving southwest in the main figure and inset. Major strike–slip faults include Alpine/Wairau (A/Wr), Awatere (Aw), Clarence (Cl), Hope (Hp) and Porters Pass (PP) of Marlborough, plus the Dextral fault belt (DFB) of the North Island. Hikurangi Trench in the inset is marked HT. Murchison basin area thrust faults in Westland are labelled Mu. The plate convergence rate is 37 mm yr^{-1}.

[1]University of Utah, Energy and Geoscience Institute, 423 Wakara Way, Suite 300, Salt Lake City, Utah 84108, USA. [2]GNS Science, P O Box 30368, Lower Hutt, 6315 Wellington, New Zealand. [3]San Diego State University, Department of Geological Sciences, 5300 Campanile Drive, San Diego, California 92182, USA. [4]Department of Geology and Geophysics, 283 Sutton Building, University Of Utah, Utah 84112, USA. [5]Tokyo Institute of Technology, Volcanic Fluid Research Center, H84, 2-12-1 Ookayama, Meguro, Tokyo, 152-8551, Japan.

©2009 Macmillan Publishers Limited. All rights reserved

not strongly concentrated along the younger faults because they are immature[5,6].

The Marlborough strike–slip faults lie in a relatively uniform crustal column of Permian–Cretaceous greywacke (Torlesse formation) that simplifies the relationship between geophysical structure and fault zone damage and fluidization compared to, for example, the San Andreas system of the western United States[7]. Northwest of the Alpine/Wairau fault in the Westland region, both the lithologies and apparent stress regime change drastically. There resides a harder backstop of Paleozoic–Mesozoic metamorphic and plutonic rocks juxtaposed against the Torlesse formation of Marlborough (Fig. 1) by ~500 km of dextral strike–slip motion[8,9]. Modern faulting in the northwest is almost purely thrust[10], including the damaging events of surface-wave magnitude $M > 7$ of the early twentieth century in the Murchison basin. These thrusts are curiously steep, 45–70° in dip, far from well-oriented for failure under the normal conditions of maximal compressive stresses being horizontal[10].

A dense transect, 200 km in length, of 67 magnetotelluric stations across the Marlborough regime was completed in 2007 to provide a detailed section view of electrical resistivity and its implications through the developing subduction system (Fig. 1). The average station spacing of about 3 km suffices to investigate the major faults as individuals in terms of their control of crustal-scale fluid flow, and in terms of correlations between deep fluidized zones, deformation and earthquakes. Data were obtained over a wave-period range of 0.004–900 s, sufficient for penetration well into the upper mantle, given the relatively high average resistivities we encountered (~1,000 ohm m). The data were transformed to a resistivity cross-section by the technique of nonlinear inversion (Methods and Supplementary Information).

The cross-section reveals several prominent features in the deep crust and upper mantle (Fig. 2). First, the relatively old (late Mesozoic) Pacific plate below its dipping seismicity is electrically resistive (>1,000 ohm m), as was observed also in magnetotelluric surveying in the central North Island[11]. Second, the western portion of the Australian plate lithosphere (Buller terrane[9]) is resistive as well, correlated possibly with the depleted roots of the Karamea batholiths. Most important in the section is a family of up to five low-resistivity (conductive) bodies (<100 ohm m) of progressively greater depth from southeast to northwest in the deep crust or upper mantle. These conductors are most probably complex aqueous fluids in the context of the Marlborough setting; likely temperatures are far from sufficient to generate such low resistivities via solid-state mechanisms[12,13].

The conductive zones imaged in Fig. 2 are interpreted to represent fluids that have migrated upward into the deep crust from source areas of progressively higher metamorphic grade along the northwest-dipping subduction zone. The fluids must possess long-range interconnection laterally of dimension comparable to their depth in order to respond significantly in the transverse magnetic mode (electric current flow across strike). A relatively direct indicator that the conductors are fluid-based is the pattern of crustal seismicity superposed in Fig. 2; with few exceptions it avoids and surrounds the conductive zones, implying that they are weak internally and that the earthquakes in part may be aftershocks triggered by fluid infiltration out of the zones[7,14,15]. High seismic P-wave attenuation, which may be caused by fluids, also correlates with lack of seismicity elsewhere on the South Island[16]. We divide the conductors into three groups (A, B and C) according to their interpreted provenances and fates, as follows (Fig. 3).

Soon after the Pacific plate is subducted under the east coast of New Zealand ('A' in Fig. 3), its pore waters and entrained soft sediments undergo primarily mechanical dewatering and clay breakdown to mica[13]. The greatest loss of water by the down-going material typically occurs at this stage, especially given that substantial Pacific plate sediments are thought to be subducted[17]. Although the pathway for much of this fluid may be back up the thrust plane, an alternate route could be almost directly overhead into the upper plate rocks. Overpressured fluids can breach the brittle–ductile transition in a fault valve mechanism forming permeable fault-fracture meshes, instigated by seismicity deep in the overlying plate[18]. The two noted conductors marked A are moderately correlated with the Hope fault zone and the projection of the Porters Pass fault zone; we therefore suggest that an evolving along-strike linkage of fluid-facilitated, fault-fracture damage zones may eventually promote formation of major strike–slip faults as in the Marlborough system.

The low-resistivity zones marked B under central Marlborough clarify a long-standing controversy on the evolution of strike–slip

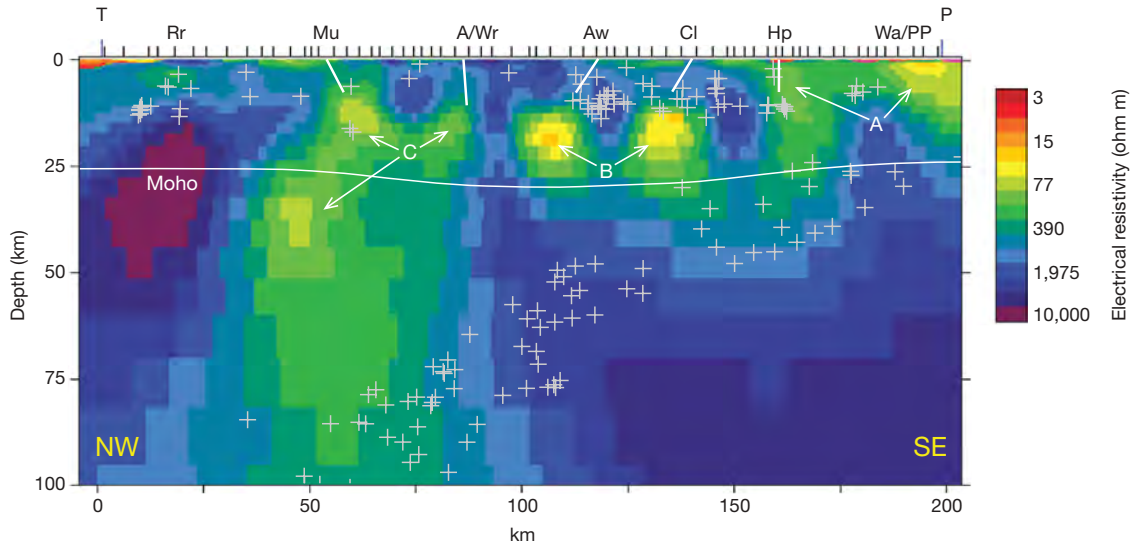


Figure 2 | Nonlinear 2D inversion model of electrical resistivity below the Marlborough–northern Westland district. The colour scale is logarithmic. Seismicity (plus symbols) is superposed for a 25-km swath along the magnetotelluric transects (provided by D. Eberhart-Phillips from temporary seismometer arrays and the New Zealand National Seismograph Network[16]). The Pacific plate lies below and to the right of the northwest-dipping deep seismicity, while the Australian plate is to the northwest. Crustal seismicity near the Awatere fault plots somewhat east of fault projection owing to a local jog in the fault strike near the transect. The curved, almost horizontal line in the 25–30-km depth range is the seismically estimated Mohorovičić discontinuity[17]. Major faults are labelled as in Fig. 1 with approximate down-dip extensions. Fault Mu is somewhat schematic, but resides near the west margin of the Murchison basin in the Buller and Owen river gorges[10]. Other physiography includes the Radiant range (Rr) and Waiau river crossing (Wa). The Tasman and Pacific oceans are labelled T and P. Major conductive zones (discussed in the text) are labelled A, B and C. The distance along *x* axis is with respect to the west coast where it intersects the profile.

©2009 Macmillan Publishers Limited. All rights reserved

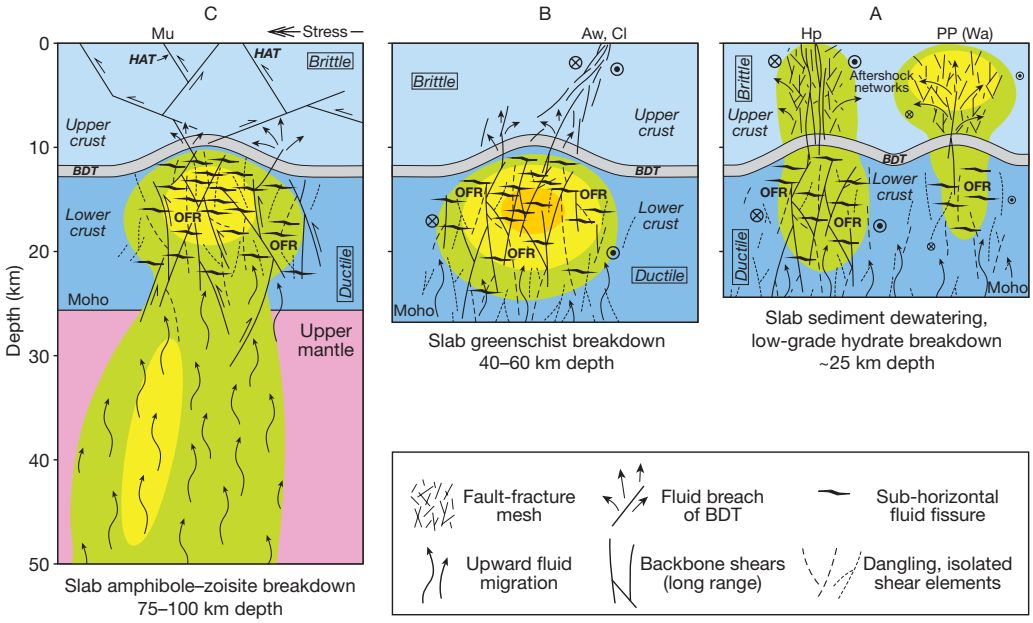


Figure 3 | Interpretative geological/fluidized states over pertinent depth ranges for labelled major conductive zones imaged below the magnetotelluric transect. BDT, brittle–ductile transition; OFR, overpressured fluid reservoir; HAT, high-angle thrust; major faults are labelled as in Figs 1 and 2. Green, yellow and orange zones are simplified representations of low resistivity fluidized zones A, B and C imaged in the inversion model of Fig. 2. The conductive zones A, B and C are described, incorporating numerous published concepts in feedback between fluid migration, accumulation, interconnection and faulting/shearing[10,18,19,23,24]. A, Fault fracture meshes with shear zone initiation. B, Narrow brittle shallow fault with broad ductile deep shearing. C, Dominantly compressional, high-angle fluidized thrusts. Undulation in the BDT is expected from fluid and strain weakening, perhaps triggered by local compositional heterogeneity[24]. Moho, Mohorovičić discontinuity.

faults with depth in the lithosphere[19,20] ('B' in Fig. 3). These broad zones lie at the down-dip projections of the Clarence and Awatere faults (Fig. 2)[21,22] and confirm the view that narrow, strike–slip deformation in the brittle upper crust evolves through stress dissipation to wider strike–slip shear in the ductile deep crust. The weak low resistivity joining the deeper reaches of the B zones with the basal Hope fault conductor is consistent with the shear zones further broadening or merging with depth into a sub-horizon of deformation[19]. Fluid reduces rock strength to very low values through diffusion creep, contributing to stress dissipation[23].

Fluids for the B zones are sourced almost continuously along the deepening subduction plate interface through metamorphic dehydration reactions in greenschist and lower amphibolite facies[13] and rise buoyantly to the crust. Transpressional deformation in ductile rocks helps to establish long-range hydraulic interconnection and thus electrical conduction in through-going backbone shear elements[23,24]. Moreover, regional compression creates subhorizontal fissuring that can hold fluids in a quasi-stable state, with occasional upward drainage induced by fault-valve or other tectonic events[24,25]. We suggest that the B zones represent a real-time view of the globally relevant process described as producing mesozonal gold mineralization, in which tapped fluids suddenly enter disequilibrium conditions[23,24]. That the shallower Awatere and Clarence faults in the brittle crust are not conductive may indicate simply that fluidization is intermittent or diffuse.

The fluid, stress and structural ramifications of conductivity zones marked C of Westland in Fig. 2 are quite different from the zones of interior Marlborough. They carry the profound implication that deep subduction fluids can induce crustal seismogenesis ('C' in Fig. 3). The major Murchison basin earthquakes are remarkable in occurring on high-dip-angle (>45°) planes that are not favourably disposed for activation in dry conditions, apparently necessitating episodic imposition of high fluid pressures to allow failure[10]. Low resistivity structure suggests that failure-enabling fluids may originate at depths of the order of 100 km near the subduction interface, presumably through amphibole–zoisite breakdown[13]. Also in the upper 20 km, a lobe of low resistivity in zone C projects southeastward towards the Alpine/Wairau fault, skirting below resistive median batholith rocks (Fig. 2). Frequent fluid input and the maturity of this long-offset fault may negate strike–slip stresses, causing a notable lack of local seismicity[16]. Following previous assessments[25], the required fluid content for all the deep conductors resolved is not high; minimal fluid contents are <0.1 vol.%, assuming high salinities and efficient fracture interconnection, although true contents would presumably be greater.

If the Hikurangi subduction zone continues to propagate southwestward[14], today's compression across Westland could ultimately give way to extension and magmatism as seen in the Taupo volcanic zone of the North Island. Magnetotelluric surveying at Taupo revealed vastly more pronounced low resistivity in the deep crust due to magmatic intrusion and fluid exsolution[11]. Fully developed subduction systems elsewhere in the world, such as the US Cascadia[26] and South America[27], show similar features. In studies with sufficient aperture and long period measurements, strong conductors are also seen in the mantle wedge, either below the arc or in the extensional back-arc regimes, and have been associated with melt-source regions.

By comparison, the central South Island southwest of Marlborough (Fig. 1) exhibits an even earlier stage of compressional development with only a crustal root formed but no subduction plane[6]. Plate convergence manifests as the frontal thrust of the Alpine fault to the northwest, rapid uplift and erosion towards the Main Divide, and a series of back thrusts to the southeast, so that large-scale detachments lie within the crust. A previous magnetotelluric transect across the central South Island[25,28] showed a striking U-shaped conductor under the entire central orogen, implying fluids generated through crustal thickening, strain and prograde metamorphism, and with long-range interconnection promoted by shearing. The elevated pore pressures created thereby are predicted in geodynamic models to augment dilation in the middle and upper crust above, especially near the Main Divide, and this is confirmed by observed mesothermal gold deposition and high-angle, veined back-shears[29]. Global positioning system geodetic data across the orogen are fitted well by crustal block models that consider the magnetotelluric conductor as a zone of concentrated

©2009 Macmillan Publishers Limited. All rights reserved

shearing[30]. Observed seismicity again supports the concept of the conductor being a weak zone, with overlying resistive rocks representing strong crust (Supplementary Information). In the absence of subduction, central South Island fluids inferred from the conductivity may thus be considered internally generated. This is in contrast to Marlborough where fluids are primarily externally generated at depth in the subduction zone before migration to their positions in the crust.

METHODS SUMMARY

Magnetotelluric measurement is a passive geophysical technique that uses naturally occurring, planar electromagnetic fields as sources for probing the Earth's resistivity (or its inverse, conductivity) (see Methods). With ultra-remote referencing and noise-avoidance steps, and with 2–3-day recording times for each site, data errors in our project were typically ~2% in magnetotelluric impedance over most of the period range. The geology and tectonics of Marlborough support a northeast–southwest strike, and this was confirmed by the principal axes of the magnetotelluric response. Nevertheless, no natural setting is purely two-dimensional (2D) and we constructed our model resistivity section by emphasizing the nominal transverse magnetic mode data subset (current flow across geological strike) together with the vertical magnetic field (current flow along strike). These quantities are relatively robust to variations in structure along the presumed strike, whereas the transverse electric mode (electric field along strike) is de-emphasized owing to this concern. Models of subsurface electrical resistivity were derived through standard nonlinear regularized inversion methods, somewhat analogous to seismic tomography. Model resolution tests were made by observing departures from soft a priori constraints and by fixing portions of the model to alternate values and assessing the misfit after re-inversion.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 February; accepted 8 June 2009.

1. Stern, R. J. Subduction initiation: spontaneous and induced. *Earth Planet. Sci. Lett.* **226,** 275–292 (2004).
2. Schellart, W. P., Freeman, J., Stegman, D. R., Moresi, L. & May, D. Evolution and diversity of subduction zones controlled by slab width. *Nature* **446,** 308–311 (2007).
3. Walcott, R. I. Models of oblique compression: Late Cenozoic tectonics of the South Island of New Zealand. *Rev. Geophys.* **36,** 1–26 (1998).
4. Furlong, K. P. in *Exhumation Associated with Continental Strike-Slip Fault Systems* (eds Till, A. B., Roeske, S. M., Sample, J. C. & Foster, D. A.) Geol. Soc. Am. Spec. Pap. 434, 1–14 (GSA, 2008).
5. Wesnousky, S. G. Seismicity as a function of cumulative geologic offset: some observations from southern California. *Bull. Seismol. Soc. Am.* **80,** 1374–1381 (1990).
6. Davey, F. J. *et al.* in *A Continental Boundary: Tectonics at South Island, New Zealand* (eds Okaya, D., Stern, T. & Davey, F.) Geophys. Monogr. 175, 47–73 (American Geophysical Union, 2007).
7. Becken, M. *et al.* A deep crustal fluid channel into the San Andreas Fault system near Parkfield, California. *Geophys. J. Int.* **173,** 718–732 (2008).
8. Mortimer, N. New Zealand's geological foundations. *Gondwana Res.* **7,** 261–272 (2004).
9. Jongens, R. Structure of the Buller and Takaka terrane rocks adjacent to the Anatoki fault, northwest Nelson, New Zealand. *NZ J. Geol. Geophys.* **49,** 443–461 (2006).
10. Ghisetti, F. C. & Sibson, R. H. Accommodation of compressional inversion in north-western South Island (New Zealand): old faults versus new? *J. Struct. Geol.* **28,** 1994–2010 (2007).
11. Heise, W. *et al.* Melt distribution beneath a young continental rift: the Taupo Volcanic Zone, New Zealand. *Geophys. Res. Lett.* **34,** L14313, doi:10.1029/2007GL029629 (2007).
12. Jones, A. G. Imaging the continental upper mantle using electromagnetic methods. *Lithos* **48,** 57–80 (1999).
13. Peacock, S. M. in *Inside the Subduction Factory* (ed. Eiler, J.) Am. Geophys. Monogr. 138, 7–22 (American Geophysical Union, 2003).
14. Reyners, M., Eberhart-Phillips, D. & Stuart, G. The role of fluids in lower-crustal earthquakes near continental rifts. *Nature* **446,** 1075–1079 (2007).
15. Ogawa, Y. & Honkura, Y. Mid-crustal electrical conductors and their correlations to seismicity and deformation at Itoigawa-Shizuoka tectonic line, central Japan. *Earth Planets Space* **56,** 1285–1292 (2004).
16. Eberhart-Phillips, D., Chadwick, M. & Bannister, S. Three-dimensional attenuation structure of central and southern South Island, New Zealand, from local earthquakes. *J. Geophys. Res.* **113,** B05308, doi:10.1029/2007JB005359 (2008).
17. Eberhart-Phillips, D. & Henderson, M. C. Including anisotropy in 3-D velocity inversion and application to Marlborough, New Zealand. *Geophys. J. Int.* **156,** 237–254 (2004).
18. Cox, S. F. Coupling between deformation, fluid pressures, and fluid flow in ore-producing hydrothermal systems at depth in the crust. *Soc. Econ. Geol.* 100th Anniv. Vol., 39–75 (2005).
19. Bourne, S. J., England, P. C. & Parson, B. The motion of crustal blocks driven by flow in the lower lithosphere and implications for slip rates of continental strike-slip faults. *Nature* **391,** 655–659 (1997).
20. Wilson, C. K., Jones, C. H., Molnar, P., Sheehan, A. F. & Boyd, O. S. Distributed deformation in the lower crust and upper mantle beneath a continental strike-slip fault zone: Marlborough fault system, South Island, New Zealand. *Geology* **32,** 837–840 (2004).
21. Nicol, A. & Van Dissen, R. Up-dip partitioning of displacement components on the oblique-slip Clarence fault, New Zealand. *J. Struct. Geol.* **24,** 1521–1535 (2002).
22. Mason, D. P. M., Little, T. A. & Van Dissen, R. J. Rates of active faulting during late Quaternary fluvial terrace formation at Saxton River, Awatere fault, New Zealand. *Geol. Soc. Am. Bull.* **118,** 1431–1466 (2006).
23. Cox, S. F. in *Fractures, Fluid Flow and Mineralization* (ed. McCaffrey, K. J. W., Lonergan, L. & Wilkinson, J. J.) Geol. Soc. Lond. Spec. Pub. 155, 123–140 (GSL, 1999).
24. Sibson, R. H. in *Deformation of the Continental Crust: the Legacy of Mike Coward* Geol. Soc. Lond. Spec. Publ. 272, 519–532 (GSL, 2007).
25. Wannamaker, P. E. *et al.* Fluid generation and pathways beneath an active compressional orogen, the New Zealand Southern Alps, inferred from magnetotelluric data. *J. Geophys. Res.* **107,** doi:10.1029/2001JB000186 (2002).
26. Patro, P. K. & Egbert, G. D. Regional conductivity structure of Cascadia: preliminary results from 3D inversion of USArray transportable array magnetotelluric data. *Geophys. Res. Lett.* **35,** doi:10.1029/2008GL035326 (2008).
27. Brasse, H. & Eydam, D. Electrical conductivity beneath the Bolivian orocline and its relation to subduction processes at the South American continental margin. *J. Geophys. Res.* **113,** doi:10.1029/2007JB005142 (2008).
28. Jiracek, G. R., Gonzalez, V. M., Caldwell, T. G., Wannamaker, P. E. & Kilb, D. in *A Continental Boundary: Tectonics at South Island, New Zealand* (eds Okaya, D., Stern, T. & Davey, F.) Geophys. Monogr. Ser. 175, 75–94 (Am. Geophys. Union, 2007).
29. Upton, P. & Koons, P. O. in *A Continental Boundary: Tectonics at South Island, New Zealand* (eds Okaya, D., Stern, T. & Davey, F.) Geophys. Monogr. Ser. 175, 253–270 (Am. Geophys. Union, 2007).
30. Beavan, J., Ellis, S. & Wallace, L. in *A Continental Boundary: Tectonics at South Island, New Zealand* (eds Okaya, D., Stern, T. & Davey, F.) Geophys. Monogr. Ser. 175, 75–94 (Am. Geophys. Union, 2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by the Geophysics program of the US National Science Foundation (grant EAR0440050), and the Plate Boundary program of the New Zealand Foundation for Research, Science and Technology. W. Hales of Alpine Springs Helicopters provided an airborne transport service to remote locations of the Marlborough and Westland regions. We thank B. Freer, C. Davis and P. Thorton of the New Zealand Department of Conservation, and numerous private landholders, for permission to access site locations. Additional field assistance was given by students M. Burgess and P. Winther. Many discussions were held with D. Eberhart-Phillips, R. Sibson and P. Upton. Illustrations were finalized by D. Jensen.

Author Contributions P.E.W., T.G.C. and G.R.J. designed the experiment. The home institutions of T.G.C. and Y.O. supplied the instrumentation. T.G.C. and G.J.H. reduced the observed time series. V.M. derived the induction vectors. All authors were essential to the success of the field campaign and contributed to the interpretation and the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.E.W. (pewanna@egi.utah.edu).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

General reviews of the magnetotelluric method are widely published[31,32]. The 67 magnetotelluric stations of the Marlborough project were recorded using commercially manufactured systems by Phoenix Geophysics Ltd (V-2000 model), owned by GNS Science and the Tokyo Institute of Technology. The systems use high-moment induction coils to measure the three components of magnetic field with noise floors of the order of 10^{-4} nT at middle periods. Electric fields were measured using 100-m-long orthogonal bipoles oriented along, and normal to, geomagnetic north and contacting the ground at their endpoints with Pb–$PbCl_2$ non-polarizing electrodes of standard design, buried typically ~30 cm deep[33]. Recording times were 2–3 days.

Stations were usually far from urban areas and typical man-made electromagnetic noise sources included electrified livestock fences and 50 Hz power supply lines. However, also running along much of the Clarence fault is the gigawatt-scale, Benmore–Wellington direct-current power transmission line with strong non-magnetotelluric current fluctuations that are especially damaging to the magnetotelluric data in the 1–10-s period band in which natural signals are weak. This structure affected most sites in the survey and necessitated the addition of a distant remote reference, located ~400 km to the northeast on the North Island to lie outside its influence[34].

Even so, the Benmore–Wellington line noise was powerful enough not to be feasibly cancelled by standard referencing over 2–3 days for sites within ~25 km on either side of it. Our campaigns were thus coordinated to include three maintenance downtimes for this line of 36–48 h, during which we mustered up to ten simultaneous magnetotelluric sites in order to achieve clean records. All time series were subsequently processed using robust remote reference methods[35] with total recording time subdivided into 2-h segments the quality of which could be assessed individually. Average sounding recording times of 2–3 days generated data points with errors typically $<1\%$ of a $\log_{10}$ unit in apparent resistivity and 0.66° in impedance phase (the floor used in Fig. 3 of the Supplemental Information). Vertical magnetic field (tipper) data quality was good but degraded somewhat compared to impedance by unavoidable wind and thermal noise.

Section views of resistivity from a single profile of magnetotelluric stations are meaningful if the data behave in a 2D manner to a good approximation. For a purely 2D case, magnetotelluric data separate into two independent modes of response. These are the transverse magnetic mode, corresponding to electric current flow across strike, and the transverse electric mode, for current flow along strike[31]. To confirm that deep crustal and upper mantle resistivity trends match our modelling assumptions well, we derived the impedance phase tensor ellipses for all stations for periods from 2.67 to 170 s, corresponding to middle-crustal to upper-mantle depths[36] (Supplementary Information). The axes align reasonably well along and across apparent geological strike and principal phase values reveal mid-crustal conductors associated with the major fault zones, including below the Westland thrust regime of the Murchison basin. Induction vectors from the vertical magnetic field are small at short periods with increasing coast effects at long periods, owing to the flanking conductive seawater (Supplementary Information). They include influence by the seawater passage through Cook strait to the northeast, as seen in previous induction vector surveys. Such off-profile structural effects have typically been treated as separable for 2D modelling.

The 2D nonlinear inversion algorithm applied was developed in-house using the finite-element method and a Gauss–Newton parameter step (Supplementary Information). Model solution is stabilized by damping roughness relative to an a priori model. The a priori structure was a one-dimensional variation derived from an integral of the data with the addition of the known geometry and resistivity of the South Island's flanking seawater and sedimentary bodies, using published cross-sections and wellbore logs. Various roughness damping weights were tried to ensure the persistence of important structures and key elements were tested by imposing alternate model properties and assessing the resultant misfit.

Model construction emphasized the transverse magnetic mode impedance (cross-strike current flow) and the vertical magnetic field (along-strike current flow) as three-dimensional simulations and experience has shown them to be relatively robust to structural changes along strike[37]. These two data subsets jointly and independently support deep crustal conductive structures associated with the main fault zones of the Marlborough region (Supplementary Information). In addition, the resolution of large-scale conductive fault structures is enhanced in the transverse magnetic mode by vertical current flow effects, whereby electric currents induced in regional lateral conductors such as the seawater or marine sediments are channelled down major fault zones to complete a current return path[25,32,38]. The vertical magnetic field coast effect inland is diagnostically suppressed by the interior Marlborough conductors at middle to longer periods. To first order, the transverse electric impedance is compatible with the model as well, although local and intermediate-scale three-dimensional effects substantially distort the response.

31. Vozoff, K. in *Electromagnetic Methods in Applied Geophysics* (ed. Nabighian, M. N.)2B, 641–711 (Soc. Explor. Geophys., 1991).
32. Booker, J. R., Favetto, A. & Pomposiello, M. C. Low electrical resistivity associated with plunging of the Nazca flat slab beneath Argentina. *Nature* **429,** 399–404 (2004).
33. Petiau, G. & Dupis, A. Noise, temperature coefficient, and long time stability of electrodes for telluric observations. *Geophys. Prospect.* **28,** 792–804 (1980).
34. Wannamaker, P. E. *et al.* Magnetotelluric surveying and monitoring at the Coso geothermal area, California, in support of the Enhanced Geothermal Systems concept: survey parameters and initial results. *Proc. Workshop Geothermal Reservoir Engr.* SGP-TR-175, 1–8 (Stanford University, 2004).
35. Jones, A. G., Chave, A. D., Egbert, G., Auld, D. & Bahr, K. A comparison of techniques for magnetotelluric response function estimation. *J. Geophys. Res.* **94,** 14201–14213 (1989).
36. Caldwell, T. G., Bibby, H. M. & Brown, C. The magnetotelluric phase tensor. *Geophys. J. Int.* **158,** 457–469 (2004).
37. Wannamaker, P. E. in *Three-Dimensional Electromagnetics* (ed Oristaglio, M. & Spies, B.) Geophys. Devel. Ser. 7, 349–374 (Soc. Explor. Geophys., 1999).
38. Wannamaker, P. E. *et al.* Lithospheric dismemberment and magmatic processes of the Great Basin-Colorado Plateau transition, Utah, implied from magnetotellurics. *Geochem. Geophys. Geosyst.* **9,** Q05019, doi:10.1029/2007GC001886 (2008).

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

New flutes document the earliest musical tradition in southwestern Germany

Nicholas J. Conard[1], Maria Malina[2] & Susanne C. Münzel[3]

Considerable debate surrounds claims for early evidence of music in the archaeological record[1–5]. Researchers universally accept the existence of complex musical instruments as an indication of fully modern behaviour and advanced symbolic communication[1] but, owing to the scarcity of finds, the archaeological record of the evolution and spread of music remains incomplete. Although arguments have been made for Neanderthal musical traditions and the presence of musical instruments in Middle Palaeolithic assemblages, concrete evidence to support these claims is lacking[1–4]. Here we report the discovery of bone and ivory flutes from the early Aurignacian period of southwestern Germany. These finds demonstrate the presence of a well-established musical tradition at the time when modern humans colonized Europe, more than 35,000 calendar years ago. Other than the caves of the Swabian Jura, the earliest secure archaeological evidence for music comes from sites in France and Austria and post-date 30,000 years ago[6–8].

Excavations in the summer of 2008 at the sites of Hohle Fels and Vogelherd in Germany produced new evidence for Palaeolithic music in the form of the remains of one nearly complete bone flute and isolated small fragments of three ivory flutes (Figs 1 and 2). On 17 September, an excavator uncovered the most significant of these finds, the bone flute, in the basal Aurignacian deposits of archaeological horizon Vb at Hohle Fels Cave in the Ach Valley, 20 km west of Ulm. The flute was recovered in 12 pieces. The team documented 11 fragments *in situ*, and one was found during water screening. The fragments were distributed over a vertical distance of 3 cm over a horizontal area of about 10 cm by 20 cm. This flute, which we designate Hohle Fels flute 1, is by far the most complete of the musical instruments so far recovered from the caves of Swabia. The flute lay in an 8-cm-thick deposit of clayey silt with limestone clasts that directly overlies a nearly sterile deposit of red-brown, silty clay, separating the basal Aurignacian from the underlying Middle Palaeolithic deposits of archaeological horizon VI (Fig. 3).

The find density in archaeological horizon Vb is moderately high, with much flint-knapping debris, worked bone and ivory, bones of horse, reindeer, cave bear, mammoth and ibex, and burnt bone. No diagnostic human bones have been found in deposits of the Swabian Aurignacian, but we assume that modern humans produced the artefacts from the basal Aurignacian deposits shortly after their arrival in the region, following a migration up the Danube Corridor[9].

The maker of the flute carved the instrument from the radius of a griffon vulture (*Gyps fulvus*). This species has a wing span of between 230 and 265 cm and provides bones ideal for large flutes. Griffon vultures and other vultures are documented in the Upper Palaeolithic sediments of the Swabian caves with several examples identified from Gravettian and Aurignacian deposits at Geißenklösterle.

The preserved portion of flute 1 from Hohle Fels has a length of 21.8 cm and a diameter of about 8 mm (Fig. 1). Comparisons with modern specimens indicate that the unmodified radius had a length

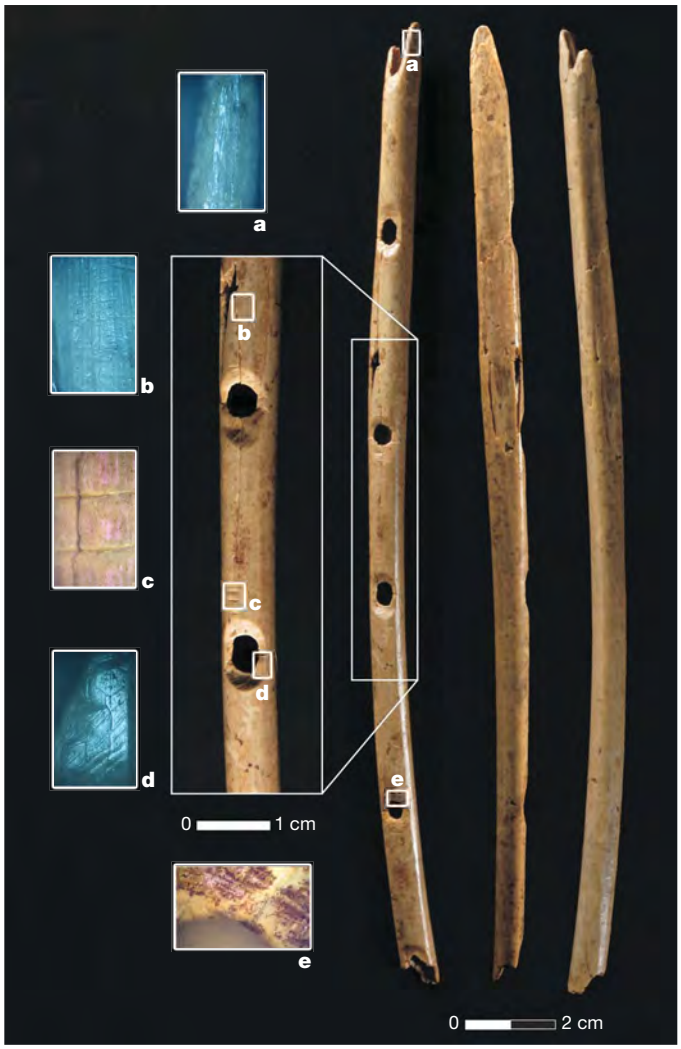


Figure 1 | Bone flute from Hohle Fels archaeological horizon Vb. Photomicrographs documenting striations and notches from manufacture and polish from use: **a**, **b**, **d**, incident-light fluorescence mode (ultraviolet- and violet-light excitation); **c**, **e**, incident light, obliquely crossed polars, λ plate. The photomicrographs were made with a Leica DMRX-MPV SP microscope photometer. The long axis of the micrographs is 2.8 mm long.

[1]Abteilung für Ältere Urgeschichte und Quartärökologie, Institut für Ur-und Frühgeschichte und Archäologie des Mittelalters, Universität Tübingen, Schloss Hohentübingen, 72070 Tübingen, Germany. [2]Research Project: The Role of Culture in The Early Expansions of Humans, Heidelberger Akademie der Wissenschaften, [3]Zentrum für Naturwissenschaftliche Archäologie, Universität Tübingen, Rümelinstrasse 23, 72070 Tübingen, Germany.

©2009 Macmillan Publishers Limited. All rights reserved

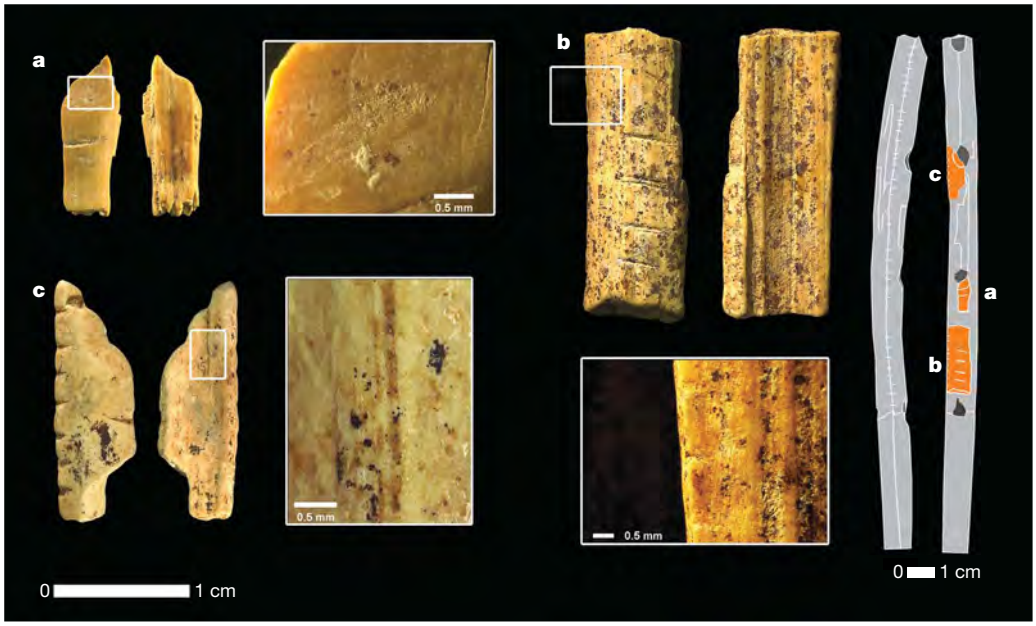


Figure 2 | Fragments of ivory flutes from Hohle Fels and Vogelherd. Photos and micrographs documenting striations and notches from manufacture and polish from use: **a**, Hohle Fels flute 2, from feature 10 of archaeological horizon Va; **b**, Hohle Fels flute 2, from archaeological horizon Vb; **c**, Vogelherd flute 2. Insets show magnified views of the boxed areas in each panel. The image on the right is a schematic superimposition of these finds on the ivory flute from Geißenklösterle; the scale is approximate. The micrographs were made using a Keyence Digital-Mikroskop VHX-500 F with reflected light.

of roughly 34 cm. The surfaces of the flute and the structure of the bone are in excellent condition and reveal many details about the manufacture of the flute. The flute has five finger holes. The maker carved two deep, V-shaped notches into one end of the instrument, presumably to form the proximal end of the flute, into which the musician blew. This end of the flute corresponds to the proximal end of the radius. The other end of the flute is broken in the middle of the most distal of the five finger holes. Several centimetres of the flute are missing from this end. As many as four very fine lines were incised near the finger holes. These precisely carved markings probably reflect measurements used to indicate where the finger holes were to be carved using chipped-stone tools. Only the partly preserved, and most distal, of the five finger holes lacks such markings.

We have not yet been able to produce a replica of the flute, but it is possible that the flute was played by blowing directly into the proximal end without using a mouthpiece. The smaller, three-holed bone flute, made from the radius of a swan, that was recovered from the Aurignacian deposits of archaeological horizon II at the nearby cave of Geißenklösterle can be played by blowing obliquely into its proximal end to produce four basic notes[10–13]. Three additional overtones can be produced by blowing more sharply into the flute. Given that the three-holed flute from Geißenklösterle produces a range of notes comparable to many modern kinds of flute, we expect flute 1 from Hohle Fels to provide a comparable, or perhaps greater, range of notes and musical possibilities[14]. The larger diameter of the bone flute from Hohle Fels would have made its tone deeper than that of the bone flute from Geißenklösterle, and closer to that documented experimentally from a reconstruction of the ivory flute from archaeological horizon II at Geißenklösterle[15].

The 2008 excavations at Hohle Fels also recovered two small fragments of what are probably two ivory flutes from the basal Aurignacian (Fig. 2). One fragment, designated flute 2, is 11.7 mm long, 4.2 mm wide and 1.7 mm thick, and comes from feature 10 at the base of archaeological horizon Va, directly overlying archaeological horizon Vb. The other fragment, which has been designated flute 3, has dimensions of 21.1 mm by 7.6 mm by 2.5 mm and originates from the lowest Aurignacian unit of archaeological horizon Vb. Crew members recovered both finds during water screening. Finds from water screening can be localized to a 10-l volume corresponding to a roughly 3-cm thickness of sediment over an area of 0.25 m^2. Both pieces of worked ivory have been hollowed out and preserve striations from their manufacture on their internal and external surfaces. The fragment of flute 2 includes a portion of a finger hole. The fragment of flute 3 preserves a series of incised lines on the convex outer surface and nine small notches along one of the edges. The greater thickness and larger dimensions of flute 3 relative to flute 2 indicate that the two finds are probably not from the same instrument.

Excavators at Vogelherd in the Lone Valley, 25 km northwest of Ulm, recovered an isolated fragment of another ivory flute, which we designate Vogelherd flute 2, while sorting water-screening samples in the summer of 2008 (Fig. 2). In 2005, we recovered three fragments of a bone flute at Vogelherd, designated Vogelherd flute 1 (ref. 16). The new ivory fragment, which has dimensions of 17.5 mm by 5.8 mm by 1.8 mm, is more heavily weathered than those found at Hohle Fels, and has a hollowed-out form, a partly preserved finger hole and seven small notches along the edge of its long axis.

The characteristics of these three fragments of ivory are known only from the ivory flute from the upper Aurignacian deposits of Geißenklösterle archaeological horizon II (ref. 15). The technology for making an ivory flute is much more complicated than that for making a flute from a bird bone. It requires forming the rough shape along the long axis of a naturally curved piece of mammoth ivory, splitting it open at the interface of the cementum and dentine or along one of the other bedding plains in the ivory, carefully hollowing out the halves, carving the holes and then rejoining the halves of the flute with air-tight seals along the seams that connected the halves of the flute. The ivory flute from Geißenklösterle preserves dozens of finely carved notches along the edges of the two halves to facilitate binding and sealing the flute[15]. Although thousands of pieces of ivory-working debris and hundreds of ivory artefacts have been recovered from the Aurignacian deposits of Hohle Fels, Vogelherd and Geißenklösterle, only the flute fragments have the form described above and preserve a hollowed-out convex morphology, finger holes and series of notches along the edge of the long axis. Thus, we can be confident that these finds represent fragments of ivory flutes similar to the one recovered from Geißenklösterle. We

©2009 Macmillan Publishers Limited. All rights reserved

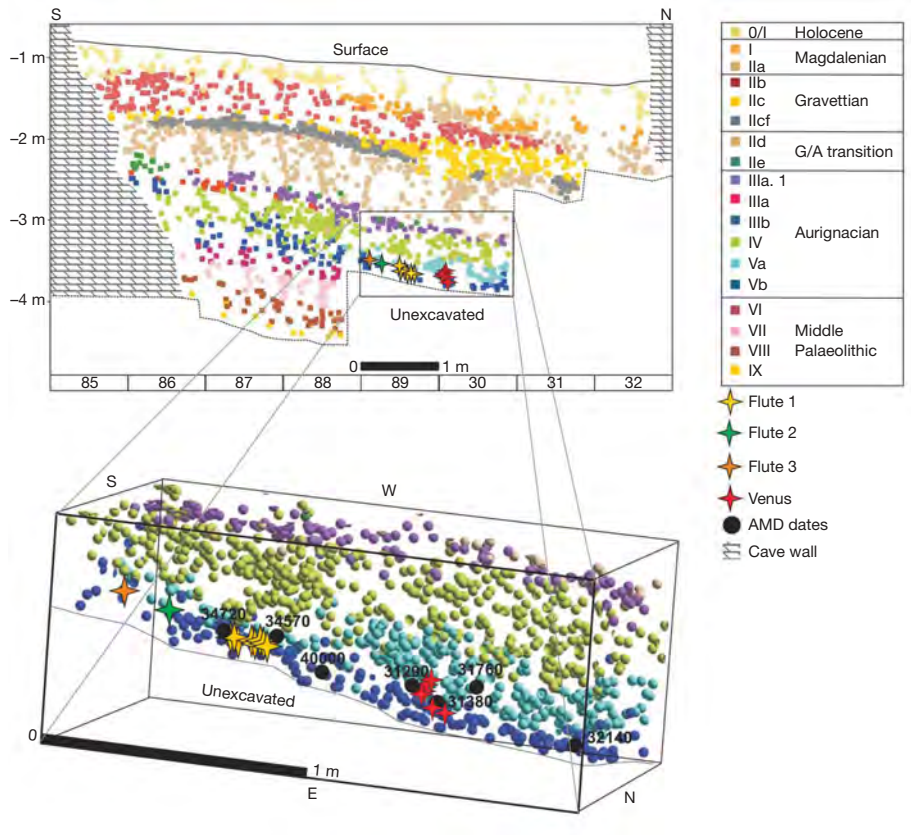


Figure 3 | The stratigraphic positions of flutes 1–3 from Hohle Fels and associated radiocarbon dates. AMS, accelerator mass spectrometry (dates in non-calibrated years before present); Venus, Venus of Hohle Fels[21].

recovered the ivory flute from Geißenklösterle in 31 small fragments. Given the tendency of delicate ivory artefacts to break into many pieces, it is not unusual to find such pieces in isolation.

The issues related to dating the Swabian Aurignacian have been a matter of considerable discussion[9,17–20]. The three flutes from Hohle Fels come from clearly documented archaeological contexts relating to the earliest Upper Palaeolithic occupation at the site (Fig. 3). Flute 2 comes from feature 10, which separates archaeological horizons Va and Vb, whereas flutes 1 and 3 come from the deepest Aurignacian stratum, archaeological horizon Vb. Several dozen radiocarbon dates from the Aurignacian of Hohle Fels have been published[21]. The Aurignacian deposits are roughly 1 m thick and include six distinct archaeological horizons and a dozen intact features. Refitting, micromorphological studies and many archaeological and geological observations indicate that deposits have not experienced significant reworking.

The ten accelerator-mass-spectrometry radiocarbon dates from archaeological-horizon-Va feature 10 and archaeological horizon Vb fall between 31 and 40 kyr ago. The radiocarbon measurements were made on collagen extracted from anthropogenically modified bone or charcoal at the Oxford Radiocarbon Accelerator Unit and the Leibniz Laboratory, Kiel[21]. The bones dated were all well preserved and produced good yields of collagen. The seven dates measured in Oxford were determined using ultrafiltration; the Leibniz Laboratory follows a slightly different filtering procedure[22,23]. Many studies demonstrate that radiocarbon dates before ~30 kyr ago vary owing to factors including differential preservation, differences in sample preparation, taphonomic mixing, fluctuations in levels of atmospheric radiocarbon and imperfect reproducibility in laboratory and measuring procedures. The highly variable radiocarbon signal at Hohle Fels echoes the situation at the well-studied site of Geißenklösterle and many others dating from this period. Although there is at present no universally accepted calibration for radiocarbon dates earlier than 30 kyr ago, available calibrations and independent controls using thermoluminescence and other methods indicate that dates of approximately 32 kyr ago correspond to roughly 36 kyr ago in calibrated years[17,24]. Thus, we can be certain that the flutes from Hohle Fels pre-date 35,000 calendar years ago. The dates from the samples closest to flute 1 are some of the earliest in this series (Fig. 3).

The stratigraphic situation suggests that the flutes from the basal Aurignacian of Hohle Fels date from the initial Upper Palaeolithic settlement of the region, about 40,000 calendar years ago[9,17]. These flutes pre-date the two bone flutes and the ivory flute from the upper Aurignacian at nearby Geißenklösterle[15,21]. The fragments of an ivory

Table 1 | Aurignacian musical instruments from the Swabian Jura

Site	Flute	Archaeological horizon	Cultural group	Material	Number of pieces	Excavated	First publication
Geißenklösterle	1	II	Upper Aurignacian	Swan radius	23	1990	Ref. 10
Geißenklösterle	2	II	Upper Aurignacian	Bird bone, swan size	7	1973	Ref. 10
Geißenklösterle	3	II	Upper Aurignacian	Mammoth ivory	31	1974–1979	Ref. 15
Hohle Fels	1	Vb	Basal Aurignacian	Griffon vulture radius	12	2008	—
Hohle Fels	2	Va. 10	Basal Aurignacian	Mammoth ivory	1	2008	—
Hohle Fels	3	Vb	Basal Aurignacian	Mammoth ivory	1	2008	—
Vogelherd	1	—	Aurignacian	Bird bone	3	2005	Ref. 16
Vogelherd	2	—	Aurignacian	Mammoth ivory	1	2008	—

©2009 Macmillan Publishers Limited. All rights reserved

and a bone flute from Vogelherd are from reworked contexts, but the vast majority of the finds from the site are from secure Aurignacian contexts. Vogelherd has produced the largest Aurignacian assemblage in central Europe, and only modest amounts of material from later Upper Palaeolithic contexts are documented at the site. Numerous radiocarbon dates from the Aurignacian at Vogelherd fall between 30 and 36 kyr ago[25]. The fragment of an ivory flute and the three fragments of a bone flute discovered in 2005, like hundreds of other diagnostic finds from the Aurignacian at Vogelherd, pre-date 30 kyr ago.

Apart from that found in the caves of the Swabian Jura, there is no convincing evidence for musical instruments pre-dating 30 kyr ago. One of the 22 Upper Palaeolithic bird-bone flutes from the important site of Isturitz in the French Pyrenees could be of Aurignacian age, but it was recovered during poorly documented excavations from the early twentieth century[1,6,7]. The other flutes have been attributed to the Gravettian, Solutrean and Magdalenian deposits at the site. The only other musical instrument of roughly comparable age is a bone flute from the open-air site of Grubgraben in the Wachau of Lower Austria, which dates from about 19 kyr ago[8].

Good evidence for both bone and ivory flutes now exists from the Swabian caves of Geißenklösterle, Hohle Fels and Vogelherd. These are the only Palaeolithic cave sites in the region where systematic water screening of all archaeological deposits has been conducted. The other Aurignacian sites in the region were excavated before the 1970s without employing excavation methods appropriate for locating small, highly fragmented finds.

When the discovery of two bone flutes from the Swabian Aurignacian was reported in 1995, these finds seemed to be exceptional and unique[10]. The subsequent discovery of additional evidence for flutes from two more sites brings the total to four bone flutes and four ivory flutes (Table 1). We can now conclude that music played an important role in Aurignacian life in the Ach and Lone valleys of southwestern Germany. Most of these flutes are from archaeological contexts containing an abundance of organic and lithic artefacts, hunted fauna and burnt bone. This evidence suggests that the inhabitants of the sites played these musical instruments in diverse social and cultural contexts and that flutes were discarded with many other forms of occupational debris. In the case of archaeological horizon Vb at Hohle Fels, the location of the bone flute in a thin archaeological horizon only 70 cm away from a female figurine of similar age suggests a possible contextual link between these two finds[21].

The flutes from Hohle Fels, Vogelherd and Geißenklösterle demonstrate that a musical tradition existed in the cultural repertoire of the Aurignacian at the time modern humans settled in the upper Danube region more than 35,000 calendar years ago. The appearance of a musical tradition in the Aurignacian accompanied the development of early figurative art and numerous innovations, including a wide array of new forms of personal ornaments and new lithic and organic technologies[25,26]. The presence of music in the lives of early Upper Palaeolithic peoples did not directly produce a more effective subsistence economy and greater reproductive fitness. Viewed, however, in a broader behavioural context, early Upper Palaeolithic music could have contributed to the maintenance of larger social networks, and thereby perhaps have helped facilitate the demographic and territorial expansion of modern humans relative to culturally more conservative and demographically more isolated Neanderthal populations[27].

Received 29 March; accepted 29 May 2009.
Published online 24 June 2009.

1. d'Errico, F. *et al.* Archaeological evidence for the emergence of language, symbolism, and music—an alternative multidisciplinary perspective. *J. World Prehist.* **17,** 1–70 (2003).
2. Turk, I. (ed.) *Mousterian "Bone Flute" and Other Finds from Divje Babe I Cave Site in Slovenia* (Znanstvenoraziskovalni Center SAZU, 1997).
3. Albrecht, G., Holdermann, C.-S., Kerig, T., Lechterbech, J. & Serangeli, J. "Flöten" aus Bärenknochen – die frühsten Musikinstrumente? *Archäol. Korrespondenzblatt* **28,** 1–19 (1998).
4. d'Errico, F. Villa, P. Pinto, A. & Idarraga, R. A Middle Paleolithic origin of music? Using cave-bear bone accumulations to assess the Divje Babe I bone 'flute'. *Antiquity* **72,** 65–79 (1998).
5. Mithen, S. *The Singing Neanderthals* (Weidenfield & Nicolson, 2005).
6. Buisson, D. Les flûtes paléolithiques d'Isturitz (Pyrénées Atlantiques). *Bull. Soc. Préhist. Française* **87,** 420–433 (1990).
7. Lawson, G. & d'Errico, F. in *Studien zur Musikarchäologie III* (eds Hickmann, E., Kilmer, A. D. & Eichmann, R.) 119–142 (Orient-Archäologie 10, Maria Leihdorf, 2002).
8. Einwögerer, T. & Bernadette Käfer, B. Eine jungpaläolithische Knochenflöte aus der Station Grubgraben bei Kammern, Niederösterreich. *Archäol. Korrespondenzblatt* **28,** 21–30 (1998).
9. Conard, N. J. & Bolus, M. Radiocarbon dating the appearance of modern humans and the timing of cultural innovations in Europe: new results and new challenges. *J. Hum. Evol.* **44,** 331–371 (2003).
10. Hahn, J. & Münzel, S. Knochenflöten aus dem Aurignacien des Geißenklösterle bei Blaubeuren, Alb-Donau-Kreis. *Fundber. Baden-Württemb.* **20,** 1–12 (1995).
11. Münzel, S., Seeberger, F. & Hein, W. in *Studien zur Musikarchäologie III* (eds Hickmann, E., Kilmer, A. D. & Eichmann, R.) 107–118 (Orient-Archäologie 10, Maria Leihdorf, 2002).
12. Hein, W. & Hahn, J. in *Experimentelle Archäologie in Deutschland: Bilanz 1997* (ed. Fansa, M.) 65–73 (Isensee, 1998).
13. Seeberger, F. *Klangwelten der Altsteinzeit* [audio CD] (Urgeschichtliches Museum Blaubeuren, 2004).
14. Tarasov, N. Die älteste Flöte der Welt. *Windkanal* **1,** 6–11 (2005).
15. Conard, N. J. *et al.* Eine Mammutelfenbeinfköte aus dem Aurignacien des Geißenklösterle. *Archäol. Korrespondenzblatt* **34,** 447–462 (2004).
16. Conard, N. J. & Malina, M. Schmuck und vielleicht auch Musik am Vogelherd bei Niederstotzingen-Stetten ob Lontal, Kreis Heidenheim. *Archäol. Ausgr. Baden-Württemb.* 21–25 (2006).
17. Richter, D., Waiblinger, J., Rink, W. J. & Wagner, G. A. Thermoluminescence, electron spin resonance and ^{14}C-dating of the late middle and early upper Palaeolithic site of Geißenklösterle cave in southern Germany. *J. Archaeol. Sci.* **27,** 71–89 (2000).
18. Conard, N. J. & Bolus, M. Radiocarbon dating the late Middle Paleolithic and the Aurignacian of the Swabian Jura. *J. Hum. Evol.* **55,** 886–897 (2008).
19. Zilhão, J. & d'Errico, F. in *The Chronology of the Aurignacian and of the Transitional Technocomplexes: Dating, Stratigraphies, Cultural Implications* (eds Zilhão, J. & d'Errico, F.) 313–349 (Proc. Symp. Trabalhos de Arqueologia 33, Instituto Português de Arqueologia, 2003).
20. Jöris, O., &. Street, M. At the end of the ^{14}C time scale – the middle to upper Paleolithic record of western Eurasia. *J. Hum. Evol.* **55,** 782–802 (2008).
21. Conard, N. J. A female figurine from the basal Aurignacian of Hohle Fels Cave in southwestern Germany. *Nature* **459,** 248–252 (2009).
22. Brock, F., Bronk Ramsey, C. & Higham, T. Quality assurance of ultrafiltered bone dating. *Radiocarbon* **49,** 187–192 (2007).
23. Hüls, M. C., Grootes, P. M. & Nadeau, M.-J. How clean is ultrafiltration cleaning of bone collagen? *Radiocarbon* **49,** 193–200 (2007).
24. Weninger, B. & Jöris, O. A ^{14}C age calibration curve for the last 60 ka: the Greenland-Hulu U/Th timescale and its impact on understanding the Middle to Upper Paleolithic transition in Western Eurasia. *J. Hum. Evol.* **55,** 772–781 (2008).
25. Conard, N. J. & Bolus, M. in *Towards a Definition of the Aurignacian* (eds Bar-Yosef, O. & Zilhão, J.) 211–239 (Proc. Symp. Trabalhos de Arqueologia 45, Instituto Português de Arqueologia/American School of Prehistoric Research, 2006).
26. Bolus, M. & Conard, N. J. What can we say about the spatial-temporal distribution of early Aurignacian innovations? *Eurasian Prehist.* **5,** 19–29 (2009).
27. Conard, N. J. *et al.* in *When Neanderthals and Modern Humans Met* (ed. Conard, N. J.) 305–341 (Kerns, 2006).

Acknowledgements Many of our colleagues, including S. Bailey, H. Bocherens, M. Bolus, K. Deckers, S. Feine, H. Floss, P. Goldberg, P. Grootes, W. Hein, T. Higham, M. Hofreiter, M. Kucera, L. Moreau, L. Niven, D. Richter, S. Riehl, F. H. Smith, H.-P. Uerpmann and S. Wolf, contributed to this research. We are particularly indebted to C. E. Miller for discussions on stratigraphy, to B. Ligouis for his microscopic images of flute 1 from Hohle Fels and to P. Krönneck for identification of bird bones. This research was supported by the Deutsche Forschungsgemeinschaft, the University of Tübingen, the Heidelberger Akademie der Wissenschaften, the Landesamt für Denkmalpflege Baden-Württemberg, the Alb-Donau-Kreis, Heidelberg Cement, the Museumsgesellschaft Schelklingen and the Gesellschaft für Urgeschichte.

Author Contributions N.J.C. directs the project and wrote the paper. M.M. coordinates the excavation and laboratory work at Hohle Fels. S.C.M. conducts faunal analysis at Hohle Fels.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.J.C. (nicholas.conard@uni-tuebingen.de).

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

Advances in development reverse fertility declines

Mikko Myrskylä[1], Hans-Peter Kohler[1] & Francesco C. Billari[2]

During the twentieth century, the global population has gone through unprecedented increases in economic and social development that coincided with substantial declines in human fertility and population growth rates[1,2]. The negative association of fertility with economic and social development has therefore become one of the most solidly established and generally accepted empirical regularities in the social sciences[1–3]. As a result of this close connection between development and fertility decline, more than half of the global population now lives in regions with below-replacement fertility (less than 2.1 children per woman)[4]. In many highly developed countries, the trend towards low fertility has also been deemed irreversible[5–9]. Rapid population ageing, and in some cases the prospect of significant population decline, have therefore become a central socioeconomic concern and policy challenge[10]. Here we show, using new cross-sectional and longitudinal analyses of the total fertility rate and the human development index (HDI), a fundamental change in the well-established negative relationship between fertility and development as the global population entered the twenty-first century. Although development continues to promote fertility decline at low and medium HDI levels, our analyses show that at advanced HDI levels, further development can reverse the declining trend in fertility. The previously negative development–fertility relationship has become J-shaped, with the HDI being positively associated with fertility among highly developed countries. This reversal of fertility decline as a result of continued economic and social development has the potential to slow the rates of population ageing, thereby ameliorating the social and economic problems that have been associated with the emergence and persistence of very low fertility.

The cross-country association between total fertility rate (TFR) and HDI in 1975 and 2005 is shown in Fig. 1. In both years, the association is negative for HDI levels below the range of 0.85–0.9. As countries progressed to very advanced levels of development (HDI > 0.9) in recent years, however, the HDI–fertility relationship started to change fundamentally. As the HDI approaches levels above about 0.9, the HDI–fertility association in Fig. 1 reverses to a positive relationship: higher levels of HDI are associated with higher levels of fertility. For example, the 2005 TFR levels for countries with an HDI between 0.9 and 0.92 is on average 1.24; in contrast, the average TFR is 1.89 in countries at the highest levels of development (HDI > 0.95). These differential fertility levels at intermediate and very advanced development stages have markedly different long-term implications: the former, if prevailing in the long term in the absence of migration, indicates a halving of the population and birth cohort approximately every 40–45 years; in contrast, the latter level can sustain population replacement with relatively modest levels of in-migration[12].

Figure 2 complements the cross-sectional analysis with a longitudinal perspective that focuses on the within-country trajectories of fertility and HDI. Figure 2 includes all countries that have attained an HDI level of least 0.9 by year 2005 and for which longitudinal data from 1975 to 2005 are available (24 countries; see Supplementary Information). The TFR is shown for years 1975 and 2005 relative to the lowest TFR that was observed while a country's HDI was within the window of 0.85–0.9. The reference year is the first year in which this lowest TFR is observed. A line is then used to connect the HDI–TFR

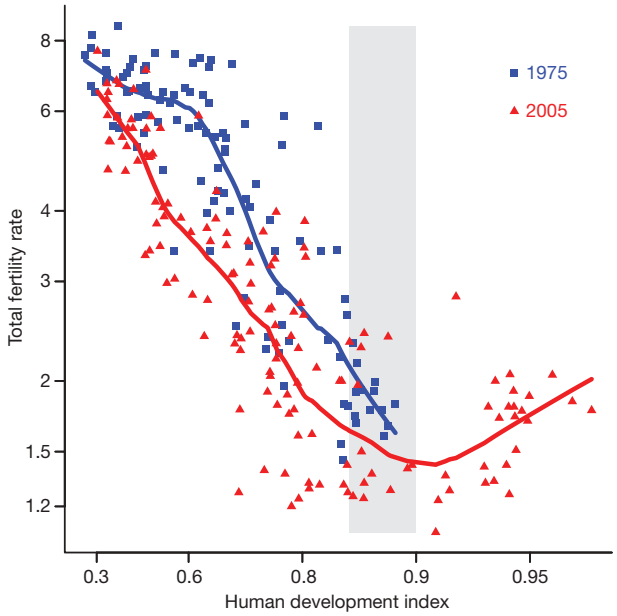


Figure 1 | Cross-sectional relationship between TFR and HDI in 1975 and 2005. The TFR reflects the number of children that would be born to a woman during her lifetime if she experienced the age-specific fertility rates observed in a calendar year. The HDI is the primary index used by the United Nations Development Programme (UNDP) to monitor and evaluate broadly defined human development, combining with equal weight indicators of a country's health conditions, living standard and human capital[11]. An HDI of 0.9 roughly corresponds to 75 years of life expectancy, a GDP per capita of 25,000 US dollars in year 2000 purchasing power parity, and a 0.95 education index (a weighted sum of standardized literacy rate and primary, secondary and tertiary level gross enrolment ratios). The 1975 data include 107 countries, with 1975 HDI levels ranging from 0.25 to 0.887, and 1975 TFR levels ranging from 1.45 to 8.5; the 2005 data include 140 countries, with 2005 HDI levels ranging from 0.3 to 0.966, and 2005 TFR levels ranging from 1.08 to 7.7. The Spearman's rank correlation between HDI and TFR in 1975 is -0.85 ($P < 0.01$); the Spearman's rank correlation between HDI and TFR in 2005 is -0.84 ($P < 0.01$) for countries with HDI < 0.85, and 0.51 ($P < 0.01$) for countries with HDI ≥ 0.9. For further details, see Supplementary Information. Countries with a 2005 HDI ≥ 0.9 include (2005 HDI in parentheses): Australia (0.966), Norway (0.961), Iceland (0.956), Ireland (0.95), Luxembourg (0.949), Sweden (0.947), Canada (0.946), Finland (0.945), France (0.945), the Netherlands (0.945), the United States (0.944), Denmark (0.943), Japan (0.943), Switzerland (0.942), Belgium (0.94), New Zealand (0.938), Spain (0.938), the United Kingdom (0.936), Austria (0.934), Italy (0.934), Israel (0.922), Greece (0.918), Germany (0.916), Slovenia (0.913) and South Korea (0.911).

[1]Population Studies Center, University of Pennsylvania, 3718 Locust Walk, Philadelphia, Pennsylvania 19104, USA. [2]DONDENA "Carlo F. Dondena" Centre for Research on Social Dynamics, Department of Decision Sciences and IGIER, Università Bocconi, via Röntgen 1, 20136 Milan, Italy.

©2009 Macmillan Publishers Limited. All rights reserved

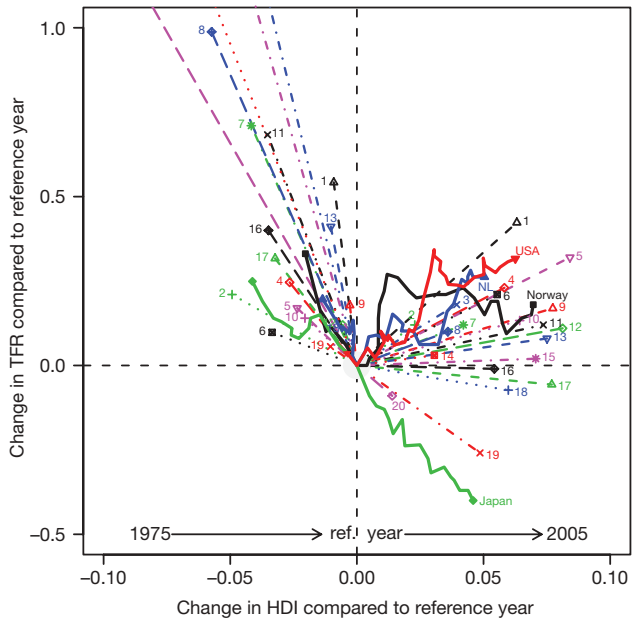


Figure 2 | Within-country time-path of the HDI–TFR relationship for all countries that attained an HDI ≥ 0.9 by 2005. The figure depicts the difference between the TFR in 1975 and 2005 compared to the lowest TFR that was observed while a country's HDI was within the 0.85–0.9 window. The (first) year in which this TFR is observed is denoted as the reference (ref.) year. For four particularly interesting and relevant countries, the United States (USA), Norway, the Netherlands (NL) and Japan, the graph shows the full path of the HDI–TFR development during the period 1975–2005. The figure includes all countries that attained an HDI ≥ 0.9 in 2005, with the exception of Slovenia for which no pre-1990 HDI time series could be constructed. For all countries, the HDI in 2005 is higher than the HDI in the reference year; for 18 of the 26 countries that attained a HDI ≥ 0.9 by 2005, the TFR in 2005 is higher than the TFR in the reference year. Countries ending in the top right quadrant in 2005 are Norway, the Netherlands, the United States, Denmark (1), Germany (2), Spain (3), Belgium (4), Luxembourg (5), Finland (6), Israel (7), Italy (8), Sweden (9), France (10), Iceland (11), the United Kingdom (12), New Zealand (13), Greece (14) and Ireland (15). Countries ending in the bottom-right quadrant in 2005 are Japan, Austria (16), Australia (17), Switzerland (18), Canada (19) and South Korea (20). See Supplementary Information for further analyses.

combinations for 1975, the reference year, and 2005. For four countries that we select as representative (Japan, the Netherlands, Norway and the United States) Fig. 2 shows the full path of the HDI–TFR changes during 1975–2005.

If the fertility–HDI relationship indeed reverses within the HDI window of 0.85–0.9, the longitudinal country-trajectories in Fig. 2 should predominantly be J-shaped. Specifically, country trajectories should begin in the top-left quadrant with relatively low HDI and high TFR levels, then pass through the circle that marks the reference year, and end in the top right quadrant where both the fertility level and the development index are higher than in the reference year. Although there are clear exceptions (as for instance Japan, Canada and South Korea), the trajectories for the large majority of countries (18 out of 24, representing 74% of the population in the 24 countries included in Fig. 2) confirm our finding of a reversal of the HDI–fertility relationship. That is, as development has progressed and these 18 countries attained an advanced HDI level of 0.9 or higher, the earlier downward trend in the total fertility rate was reversed. As a result, fertility in 2005 was higher than the minimum that was observed while a country's HDI was within the 0.85–0.9 interval. For example, US fertility reversed in 1976 (reference year) at an HDI of 0.881; the reversal in Norway occurred in 1983 at an HDI of 0.892; in Italy, the turning point occurred in 1994 at an HDI of 0.898; and in Israel, the reversal in TFR decline occurred in 1992 at an HDI of 0.880.

We confirm the graphical results of the reversal in the development–fertility relationship at advanced HDI levels by estimating a statistical model for the effect of HDI increases on fertility change (see Supplementary Information). The estimation uses panel data covering the years 1975 to 2005 for all 37 countries that had reached an HDI level of 0.85 by 2005. We use a differences-in-differences regression model with time fixed-effects[13] and a structural change in the HDI–fertility relationship at a critical HDI level that is estimated from the data. This specification controls for unobserved country characteristics and time trends, and it thus allows us to test whether the reversal in the HDI–fertility relationship documented in Fig. 2 persists after controlling for potentially confounding factors such as unobserved time-invariant country-specific factors and common time trends.

The critical HDI level at which the development–fertility association reverses from negative to positive is estimated as 0.86 (Supplementary Information). Our preferred estimates then suggest that the effect of HDI increases on fertility levels is equal to −1.59 ($P < 0.05$) for HDI levels below 0.86. The effect of HDI increases on fertility levels is estimated to be 4.07 ($P < 0.001$) for HDI levels at or above 0.86 (model 1 in Fig. 3). That is, on average an HDI increase of 0.1 results in a reduction of the TFR by 0.159 as long as countries are at development levels with HDI below 0.86; in contrast, an HDI increase of 0.05 results in an increase of the TFR by 0.204 ($= 0.05 \times 4.07$) once countries attain an advanced development stage with HDI ≥ 0.86. This fertility increase of approximately 0.2 children per woman for a 0.05 increase in HDI is sizable, and it corresponds closely to the graphical analyses presented in Fig. 2: for all countries ending in the top-right quadrant of Fig. 1, for example, TFR increased on average by 0.16 per 0.05 increase in HDI after the reference year.

The earlier finding of a positive HDI–fertility relationship at advanced HDI levels is robust even when the total fertility rate is adjusted for 'tempo effects'[7]—that is, the distortions that occur in the TFR as a result of the postponement of childbearing to later maternal ages[14]. In particular, the estimated effect of increases in HDI on the fertility level at advanced stages of development (HDI ≥ 0.86) remains positive and significant even if the tempo-adjusted TFR[15] is used as the dependent variable instead of the conventional TFR, or if the regression analyses include a further control for changes in the mean age at childbearing. Our findings are also robust with respect to alternative specifications of the statistical model that include a lagged HDI, and they are also not influenced by single data points or countries (models 2–4 in Fig. 3 and Supplementary Information).

The existence of a positive HDI–fertility relationship at advanced development stages indicates that further development has the potential to reverse earlier fertility declines once countries reach very high HDI levels. This finding has important implications in at least two domains. First, given the heterogeneity of institutional, cultural and policy contexts across developed countries, further research is required to investigate the different mechanisms that may underlie this reversal—particularly in light of exceptions such as Japan, Canada and South Korea. Specifically, an improved understanding of how improved labour-market flexibility, social security and individual welfare, gender and economic equality, human capital and social/family policies can facilitate relatively high levels of fertility in advanced societies is needed[8,16–18]. For instance, analyses on Europe show that nowadays a positive relationship is observed between fertility and indicators of innovation in family behaviour or female labour-force participation[19]. Also, at advanced levels of development, governments might explicitly address fertility decline by implementing policies that improve gender equality or the compatibility between economic success, including labour force participation, and family life[10,17,18]. Failure to answer to the challenges of development with institutions that facilitate work–family balance and gender equality might explain the exceptional pattern for rich

©2009 Macmillan Publishers Limited. All rights reserved

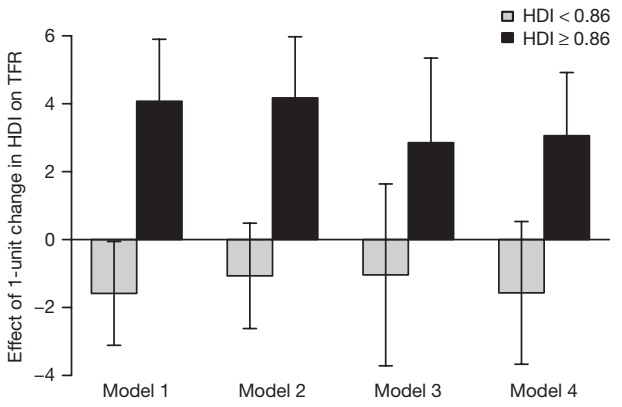


Figure 3 | Effect of 1-unit change in HDI on TFR. Estimated using differences-in-differences regression models of the HDI–TFR relationship. Analyses include time fixed-effects and allow for a structural change in the HDI–TFR relationship at a critical HDI level of 0.86 (for further details see Supplementary Table 3). Model 1 (preferred estimates): analyses include the period 1975–2005 for all countries with HDI $\geq$ 0.85 in 2005 ($n = 37$ countries; 1,051 observations). Model 2: same as model 1, except using 1-year lagged HDI as an explanatory variable ($n = 37$ countries; 1,014 observations). Model 3: same as model 1, except using tempo-adjusted TFR as a dependent variable (analyses include all countries with HDI $\geq$ 0.85 in 2005 for which a time series of the tempo-adjusted TFR is available; $n = 25$ countries; 705 observations, of which 505 include a tempo-adjusted TFR). Model 4: same as model 1, but including an additional adjustment for changes in mean age of mothers at first birth (analyses include all countries with HDI $\geq$ 0.85 in 2005 for which data on mean age at childbearing is available; $n = 26$ countries; 736 observations). The failure to identify a statistically significant negative effect of increases in HDI on the TFR at HDI levels below 0.86 in models 2–4 can be attributed to the smaller sample sizes in the alternative compared to our preferred estimates. Also, owing to the focus on the HDI–fertility relationship at advanced development levels, our regression analyses are restricted to countries that have attained a HDI of at least 0.85 by 2005, thereby excluding countries at lower levels of development for which the negative association between HDI and fertility is particularly strong (Fig. 1). Light grey bars denote moderate levels of development (HDI $<$ 0.86); dark grey bars denote advanced levels of development (HDI $\geq$ 0.86). Error bars indicate 95% confidence intervals.

eastern Asian countries that continue to be characterized by a negative HDI–fertility relationship[20].

Second, our findings are highly relevant in the debate on the future of the world's population. Whereas a decade ago Europe, North America and Japan were assumed to face very rapid population ageing and in many cases significant population declines[6,7,21], our findings provide a different outlook for the twenty-first century. As long as the most developed countries focus on increasing the well-being of their citizens, and adequate institutions are in place, the analyses in this paper suggest that increases in development are likely to reverse fertility declines—even if we cannot expect fertility to rise again above replacement levels. As a consequence, we expect countries at the most advanced development stages to face a relatively stable population size, if not an increase in total population in cases in which immigration is substantial. For countries in which immigration is a minor component of demographic change, our analyses suggest a slower population decline than is at present foreseen in official demographic forecasts. Although significant population ageing is still certain in countries at the highest development levels, its magnitude may have been exaggerated by the widely held current perception that, as social and economic development progresses, fertility is bound to fall further. Policies targeted at further increasing HDI levels in advanced societies may therefore be suitable as a general strategy to reduce demographic imbalances caused by very low fertility levels. Consistent with current scientific knowledge, our findings also support the view that progress in development contributes to lower fertility levels in countries with low to moderately high HDI levels. Moreover, countries remaining at intermediate development levels are likely to face a decline in population size because these countries have attained low TFR levels and they do not yet—and may not in the foreseeable future—benefit from the reversal of the development–fertility relationship.

Received 1 April; accepted 17 June 2009.

1. Bryant, J. Theories of fertility decline and the evidence from development indicators. *Popul. Dev. Rev.* **33,** 101–127 (2007).
2. Lee, R. D. The demographic transition: three centuries of fundamental change. *J. Econ. Perspect.* **17,** 167–190 (2003).
3. Bongaarts, J. & Watkins, S. C. Social interactions and contemporary fertility transitions. *Popul. Dev. Rev.* **22,** 639–682 (1996).
4. Wilson, C. Fertility below replacement level. *Science* **304,** 207–209 (2004).
5. Lutz, W., Sanderson, W. & Scherbov, S. The coming acceleration of global population aging. *Nature* **451,** 716–719 (2008).
6. Lutz, W., O'Neill, B. C. & Scherbov, S. Europe's population at a turning point. *Science* **299,** 1991–1992 (2003).
7. Bongaarts, J. Demographic consequences of declining fertility. *Science* **282,** 419–420 (1998).
8. Kohler, H.-P., Billari, F. C. & Ortega, J. A. The emergence of lowest-low fertility in Europe during the 1990s. *Popul. Dev. Rev.* **28,** 641–681 (2002).
9. Butler, D. The fertility riddle. *Nature* **432,** 38–39 (2004).
10. Balter, M. The baby deficit. *Science* **312,** 1894–1897 (2006).
11. United Nations Development Programme. *Statistics of the Human Development Report* (UNDP Human Development Report Office) ⟨http://hdr.undp.org/en/statistics/⟩ (29 September 2008).
12. United Nations. *Replacement Migration: Is It a Solution to Declining and Ageing Populations?* (United Nations, 2000).
13. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data.* Ch. 10 (MIT Press, 2002).
14. Sobotka, T. Is lowest-low fertility in Europe explained by the postponement of childbearing? *Popul. Dev. Rev.* **30,** 195–220 (2004).
15. Bongaarts, J. & Feeney, G. On the quantum and tempo of fertility. *Popul. Dev. Rev.* **24,** 271–291 (1998).
16. Brewster, K. L. & Rindfuss, R. R. Fertility and women's employment in industrialized nations. *Annu. Rev. Sociol.* **26,** 271–296 (2000).
17. McDonald, P. Gender equity in theories of fertility transition. *Popul. Dev. Rev.* **26,** 427–440 (2000).
18. Neyer, G. & Andersson, G. Consequences of family policies on childbearing behavior: effects or artifacts? *Popul. Dev. Rev.* **34,** 699–724 (2008).
19. Billari, F. C. & Kohler, H.-P. Patterns of low and lowest-low fertility in Europe. *Popul. Stud.* **58,** 161–176 (2004).
20. Suzuki, T. Lowest-low fertility in Korea and Japan. *J. Popul. Probl.* **59,** 1–16 (2003).
21. Lutz, W., Sanderson, W. & Scherbov, S. The end of world population growth. *Nature* **412,** 543–545 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements M.M. acknowledges support from the University of Pennsylvania, the Finnish Cultural Foundation, and the Ella and Georg Ehrnrooth foundation. H.-P.K. acknowledges the support provided by the Center for Advanced Studies at the Norwegian Academy of Science and the University of Pennsylvania. F.C.B. acknowledges support from Università Bocconi, the Italian Ministry for University and Research and the Distinguished International Scholars Program at the University of Pennsylvania. We are grateful to T. Sobotka for the provision of tempo-adjusted fertility data used in our analyses.

Author Contributions All authors contributed equally to this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.-P.K. (hpkohler@pop.upenn.edu).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 460|6 August 2009|doi:10.1038/nature08186

LETTERS

Common variants conferring risk of schizophrenia

A list of authors and their affiliations appears at the end of the paper

Schizophrenia is a complex disorder, caused by both genetic and environmental factors and their interactions. Research on pathogenesis has traditionally focused on neurotransmitter systems in the brain, particularly those involving dopamine. Schizophrenia has been considered a separate disease for over a century, but in the absence of clear biological markers, diagnosis has historically been based on signs and symptoms. A fundamental message emerging from genome-wide association studies of copy number variations (CNVs) associated with the disease is that its genetic basis does not necessarily conform to classical nosological disease boundaries. Certain CNVs confer not only high relative risk of schizophrenia but also of other psychiatric disorders[1–3]. The structural variations associated with schizophrenia can involve several genes and the phenotypic syndromes, or the 'genomic disorders', have not yet been characterized[4]. Single nucleotide polymorphism (SNP)-based genome-wide association studies with the potential to implicate individual genes in complex diseases may reveal underlying biological pathways. Here we combined SNP data from several large genome-wide scans and followed up the most significant association signals. We found significant association with several markers spanning the major histocompatibility complex (MHC) region on chromosome 6p21.3-22.1, a marker located upstream of the neurogranin gene (*NRGN*) on 11q24.2 and a marker in intron four of transcription factor 4 (*TCF4*) on 18q21.2. Our findings implicating the MHC region are consistent with an immune component to schizophrenia risk, whereas the association with *NRGN* and *TCF4* points to perturbation of pathways involved in brain development, memory and cognition.

To begin our search for sequence variants associated with schizophrenia, we performed a genome-wide scan of 2,663 schizophrenia cases and 13,498 controls from eight European locations (England, Finland (Helsinki), Finland (Kuusamo), Germany (Bonn), Germany (Munich), Iceland, Italy and Scotland; collectively called SGENE-plus) using the Illumina HumanHap300 and HumanHap550 BeadChips. In total, 314,868 SNPs meeting our quality control criteria were included in an allelic association analysis. To adjust for relatedness and potential population stratification, genomic control was applied to each study group.

None of the markers gave *P* values smaller than our genome-wide significance threshold of 0.05/314,868, or approximately 1.6×10^{-7} (see Supplementary Fig. 1 for a quantile–quantile plot and Supplementary Table 1 for markers with the smallest *P* values). Next, we combined findings from our top 1,500 markers with results for the same markers (or surrogates for them) from both the International Schizophrenia Consortium[5] (excluding the Scottish samples overlapping with samples in our study, 2,602 cases and 2,885 controls) and the European–American portion of the Molecular Genetics of Schizophrenia[6] (2,681 cases and 2,653 controls) study. Twenty-five of our top 1,500 markers (or eighteen counting very strongly correlated ($r^2 > 0.8$) markers only once) had *P* values less than 1×10^{-5} in the combined results (Supplementary Table 2). These top markers were followed up in as many as 4,999 cases and 15,555 controls from four sets of additional samples from Europe (set 1, 715 cases and 3,634 controls from the Netherlands; set 2, 3,330 cases and 6,892 controls from Denmark (Aarhus), Denmark (Copenhagen), Germany (Bonn), Germany (Munich), Hungary, the Netherlands, Norway, Russia and Sweden; set 3, 287 cases and 3,987 controls from Finland; set 4, 667 cases and 1,042 controls from Spain (Santiago) and Spain (Valencia)) (Supplementary Table 3).

Three markers, all in the extended MHC region on the short arm of chromosome 6, showed genome-wide significance in the combination of SGENE-plus and the follow-up samples described above (Table 1). In addition, four other markers—two in the MHC region, one at 11q24.2 and one at 18q21.2—showed genome-wide significance when results from the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia study were included (Table 1).

In the MHC region on chromosome 6p21.3-22.1, the five genome-wide significant markers (*P* ranging from 1.1×10^{-9} to 1.4×10^{-12} in all samples combined) have risk alleles with average control frequencies between 78% and 92% (Table 1). Combined odds ratios (ORs) for the markers range from 1.15 to 1.24 (Table 1) with no significant heterogeneity between the study groups ($P > 0.25$, Supplementary Table 4). For all of the markers, the multiplicative model for risk provides an adequate fit ($P > 0.62$).

Despite spanning about five megabases (Mb), the five chromosome 6p markers cover only about 1.4 centimorgans (cM) and substantial linkage disequilibrium exists between them (Supplementary Table 5). The association of rs6932590 (the most significant marker), however, cannot account for all of the association of the four remaining markers (Supplementary Table 6). Most notably, conditional on rs6932590, rs3131296 has an association *P* value of 3.4×10^{-6}, indicating that rs3131296 may be capturing a second susceptibility variant or that both rs6932590 and rs3131296 are correlated with a third, higher risk, variant not examined here.

To examine association of the genome-wide significant SNPs in the 5-Mb region on 6p21.3-22.1 with classical human leukocyte antigen (HLA) alleles, long-range phasing haplotypes[7] tagging the major alleles at the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* loci in Icelanders were used. Only rs3131296 shows substantial ($r^2 > 0.5$) correlation with any of the classical HLA alleles tested; this marker has an r^2 of 0.86 with *DRB1*03* and an r^2 of 0.81 with *HLA-B*08*. Simplified tags for these two classical alleles, appropriate for the European samples of SGENE-plus, had effects that were not statistically distinguishable from the effect of rs3131296. In the case of both *DRB1*03* and *HLA-B*08*, the classical HLA allele is paired with the protective allele of rs3131296, making the results described here consistent with the under-transmission of *DRB1*03* to schizophrenic offspring reported previously[8].

Many autoimmune and infectious diseases have been associated with *DRB1*03* and, indeed, inspection of top MHC region SNPs from recent genome-wide association scans of three of these—type I diabetes[9], coeliac disease[10] and systemic lupus erythematosus[11]—reveals, for each disease, SNPs having a HapMap CEU (Utah residents with ancestry from northern and western Europe) r^2 of at least 0.73

©2009 Macmillan Publishers Limited. All rights reserved

Table 1 | Genome-wide significant association of seven markers with schizophrenia

Chromosome/ megabases	SNP[allele]	Frequency	SGENE-plus* (2,663 cases; 13,498 controls)		Follow-up (4,999 cases; 15,555 controls)		SGENE-plus + follow-up (7,662 cases; 29,053 controls)		SGENE-plus + follow-up + ISC + MGS (12,945 cases; 34,591 controls)		Region/ neighbouring gene
			OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	
6/27.2	rs6913660[C]†☆	0.85	1.22 (1.10, 1.36)	0.00023	1.11 (1.04, 1.19)	0.0021	1.14 (1.08, 1.21)	4.7×10^{-6}	1.15 (1.10, 1.21)	1.1×10^{-9}	MHC/ *HIST1H2BJ*
6/27.3	rs13219354[T]‡☆	0.90	1.25 (1.11, 1.42)	0.00043	1.19 (1.08, 1.30)	0.00022	1.21 (1.12, 1.30)	4.4×10^{-7}	1.20 (1.14, 1.27)	1.3×10^{-10}	MHC/ *PRSS16*
6/27.4	rs6932590[T]§☆	0.78	1.15 (1.05, 1.26)	0.0024	1.17 (1.10, 1.25)	4.9×10^{-7}	1.17 (1.11, 1.23)	4.4×10^{-9}	1.16 (1.11, 1.21)	1.4×10^{-12}	MHC/ *PRSS16*
6/28.4	rs13211507[T]‖☆	0.92	1.24 (1.08, 1.42)	0.0027	1.27 (1.15, 1.40)	3.1×10^{-6}	1.26 (1.16, 1.36)	3.1×10^{-8}	1.24 (1.16, 1.32)	8.3×10^{-11}	MHC/ *PGBD1*
6/32.3	rs3131296[G]¶☆	0.87	1.21 (1.08, 1.36)	0.0011	1.20 (1.11, 1.30)	5.3×10^{-6}	1.21 (1.13, 1.29)	2.1×10^{-8}	1.19 (1.13, 1.25)	2.3×10^{-10}	MHC/ *NOTCH4*
11/124.1	rs12807809[T]	0.83	1.19 (1.08, 1.32)	0.00045	1.13 (1.06, 1.21)	0.00022	1.15 (1.09, 1.22)	5.0×10^{-7}	1.15 (1.10, 1.20)	2.4×10^{-9}	*NRGN*
18/51.3	rs9960767[C]#☆	0.056	1.30 (1.11, 1.51)	0.0011	1.20 (1.08, 1.33)	0.00044	1.23 (1.13, 1.34)	2.2×10^{-6}	1.23 (1.15, 1.32)	4.1×10^{-9}	*TCF4*

Allelic OR and *P* values (two-sided) based on the multiplicative model are shown. Frequency is the average allelic control frequency in SGENE-plus. Megabase is from the National Center for Biotechnology Information (NCBI) build 36. To combine the study groups within SGENE-plus and within the follow-up, the Mantel–Haenszel model was used. SGENE-plus and the follow-up sets were combined with the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia study by summing weighted summary statistics. Note that the combined analysis shown here differs from the one reported in the companion ISC and MGS papers; here we include the Aberdeen samples in SGENE-plus and not ISC, and we also incorporate additional follow-up samples. ISC, International Schizophrenia Consortium; MGS, Molecular Genetics of Schizophrenia.

* *P* values were adjusted using genomic control (see Methods).

† rs4452638 in International Schizophrenia Consortium (HapMap CEU $r^2 = 0.866$).

‡ rs4452638 in International Schizophrenia Consortium (HapMap CEU $r^2 = 1$).

§ rs3800307 in International Schizophrenia Consortium (HapMap CEU $r^2 = 0.843$).

‖ rs13214023 in International Schizophrenia Consortium (HapMap CEU $r^2 = 0.915$).

¶ rs1150753 in International Schizophrenia Consortium (HapMap CEU $r^2 = 1$).

rs10401120 in International Schizophrenia Consortium (HapMap CEU $r^2 = 0.867$).

☆ Imputed using MACH 1.0 in the Molecular Genetics of Schizophrenia study.

with rs3131296. For all of the diseases, the allele that is protective from schizophrenia is associated with the 'at-risk' allele for the autoimmune disease. This reciprocal association may, at least in part, explain the recently reported inverse association between type I diabetes and schizophrenia[12,13]. A positive association, however, has been described for coeliac disease and schizophrenia[12]; if this positive association has a genetic basis, it must be the result of associations at variants other than the one described here.

Schizophrenia patients are more likely, compared to the general population, to have been born in the winter or the spring. Although infections such as influenza and measles have been proposed as a possible mechanism for this distortion, a clear association between infectious agents and schizophrenia has not been demonstrated. The association with the MHC region reported here supports a role for infection but, as many non-immune-related genes are also found in the extended MHC region, it does not provide strong evidence. On the basis of the 3,130 schizophrenia patients for which month of birth information was available, no significant difference in the frequency of the top SNPs from the MHC region according to season of birth (winter/spring versus summer/autumn) was identified ($P > 0.29$).

The MHC region has long been postulated to harbour variants conferring risk of schizophrenia, both because of evidence for linkage in the region[14] and because of the suggested involvement of infection. Association studies of variants from the MHC region to date, however, have had modest sample sizes and therefore have lacked the power to detect effects similar to those described here.

The genome-wide significant marker ($P = 2.4 \times 10^{-9}$) at 11q24.2, rs12807809, has an average risk allele control frequency of 83% and a combined OR of 1.15 (Table 1) with no significant OR heterogeneity between the study groups ($P = 0.74$, Supplementary Table 4). The multiplicative model provides an adequate fit ($P = 0.18$). This marker is 3,457 bases upstream of neurogranin (*NRGN*). *NRGN* has previously been reported as associated in males with schizophrenia in a small Portuguese series[15], although the associated SNP in that paper, rs7113041, is not closely correlated with the SNP reported here (HapMap CEU $r^2 = 0.11$). Furthermore, reduced NRGN immunoreactivity has been observed in prefrontal areas 9 and 32 of post-mortem schizophrenia brains[16]. *NRGN* is expressed exclusively in brain, especially in dendritic spines, with expression directly controlled by thyroid hormone[17]. It is therefore possible that the psychotic and cognitive features associated with thyroid dysfunction may, in part, be mediated through dysregulation of *NRGN* gene expression.

NRGN encodes a postsynaptic protein kinase substrate that binds calmodulin (CaM) in the absence of calcium; it is abundantly expressed in brain regions important for cognitive functions, and is especially enriched in CA1 pyramidal neurons in the hippocampus[18]. The main function of NRGN may be to act as a CaM reservoir, regulating its availability in the postsynaptic compartment. Glutamate stimulation of *N*-methyl-D-aspartate (NMDA) receptors results in Ca^{2+} influx to the neuron, NRGN oxidation and release of CaM[19]. The consequent activation of postsynaptic calcium/calmodulin-dependent protein kinase II (CaMKII) by CaM results in a sustained strengthening of synaptic connections; conversely, CaM activation of calcineurin (PP2B) weakens these connections. CaMKII has a major role in mediating the NMDA-receptor signalling involved in synaptic plasticity and formation of associative memories in the brain[20]. Impaired memory function is thought to be a core feature of the pathophysiology of schizophrenia[21], especially affecting short term memory where CAMKII is thought to have a major role[22]. Glutamate stimulation of NMDA receptors results in Ca^{2+} influx to the neuron and in NRGN oxidation. Altered NRGN activity may therefore mediate the effects of NMDA hypofunction implicated in the pathophysiology of schizophrenia.

On 18q21.2, marker rs9960767 has a genome-wide significant *P* value of 4.1×10^{-9} (Table 1). The risk allele control frequency is about 6% and the OR is 1.23 with no significant OR heterogeneity between the study groups ($P = 0.34$, Supplementary Table 4). The multiplicative model gives an adequate fit ($P = 0.81$). Genome scan meta-analysis of linkage studies of schizophrenia[14] ranks the 18q21.1-qter 'bin' around fifteenth in the genome. *TCF4* is essential for normal brain development[23], and mutations in the gene were recently found to be responsible for Pitt–Hopkins syndrome, an autosomal-dominant neurodevelopmental disorder characterized by severe motor and mental retardation, microcephaly, epilepsy and facial dysmorphisms[24]. The phenotype need not be as extreme as Pitt–Hopkins syndrome; a *de novo* translocation disrupting exon 4 of *TCF4* was found in an individual with problems restricted to mental retardation[25]. Thus, it seems that variants in a single gene can be

©2009 Macmillan Publishers Limited. All rights reserved

associated with a range of neuropsychiatric phenotypes including schizophrenia. This is in line with the range of phenotypes associated with CNVs recently associated with schizophrenia[1–3].

In addition to these three genome-wide significant loci, further putative susceptibility variants are highlighted by this study. For instance, in the set of 18 markers taken into follow-up studies, rs2312147 achieved a *P* value that was not far from our genome-wide significance threshold (Supplementary Table 3). Also, markers having *P* values in the combined SGENE-plus, International Schizophrenia Consortium and Molecular Genetics of Schizophrenia data set (Supplementary Table 2) that are somewhat larger than those followed up here can be investigated further. Intriguingly, these markers include rs6589386, which is located in an intergenic region upstream of *DRD2*, a candidate gene for schizophrenia.

Our findings demonstrating association of schizophrenia with markers in the MHC region are consistent with previous reports suggesting immune system involvement in schizophrenia, whereas association with *NRGN* and *TCF4* points more to perturbation of pathways involved in brain development and cognitive function, particularly memory. Impaired cognitive and memory functions are being recognized increasingly as core features of schizophrenia[21] which are poorly addressed by current medications. The three common genetic variants we describe, which predispose to schizophrenia, have the potential to be translated into targets for the development of novel medications.

METHODS SUMMARY

Subjects. SGENE-plus, the genome-wide portion of the study, included 2,663 cases and 13,498 controls from eight European locations: England, Finland (Helsinki), Finland (Kuusamo), Germany (Bonn), Germany (Munich), Iceland, Italy and Scotland. Follow-up groups comprised 4,999 cases and 15,555 controls from twelve European locations: Denmark (Aarhus), Denmark (Copenhagen), Finland, Germany (Bonn), Germany (Munich), Hungary, the Netherlands, Norway, Russia, Spain (Santiago), Spain (Valencia) and Sweden. Cases were diagnosed with schizophrenia according to DSM-IV or ICD-10 criteria (see Supplementary Methods).

Genotyping. Genome-wide genotyping was carried out at deCODE, Duke University and the University of Bonn using either HumanHap300 or HumanHap550 BeadChips (Illumina). Individual genotyping was done via Centaurus assays at deCODE and via multiplex PCR and mini-sequencing assays, followed by MALDI-TOF mass spectrometry analysis in Spain and Finland.

Statistical analysis. A likelihood procedure described previously was used for association analysis[26]. To correct for relatedness and potential population stratification, genomic control was used[27]. Within our study, samples were combined using the Mantel–Haenszel model[28]. Our results were merged with those of the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia study by computing weighted averages of *Z* scores (see Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 March; accepted 5 June 2009.
Published online 1 July 2009.

1. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).
2. Mefford, H. C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359,** 1685–1699 (2008).
3. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008).
4. Brunetti-Pierri, N. *et al.* Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature Genet.* **40,** 1466–1471 (2008).
5. The International Schizophrenia Consortium. Common polygenetic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* doi:10.1038/nature08185 (this issue).
6. Shi, J. *et al.* Common variants on chromosomes 6p22.1 are associated with schizophrenia. *Nature* doi:10.1038/nature08192 (this issue).
7. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40,** 1068–1075 (2008).
8. Li, T. *et al.* Transmission disequilibrium analysis of HLA class II DRB1, DQA1, DQB1 and DPB1 polymorphisms in schizophrenia using family trios from a Han Chinese population. *Schizophr. Res.* **49,** 73–78 (2001).
9. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
10. van Heel, D. A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nature Genet.* **39,** 827–829 (2007).
11. Harley, J. B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nature Genet.* **40,** 204–210 (2008).
12. Eaton, W. W. *et al.* Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *Am. J. Psychiatry* **163,** 521–528 (2006).
13. Juvonen, H. *et al.* Incidence of schizophrenia in a nationwide cohort of patients with type 1 diabetes mellitus. *Arch. Gen. Psychiatry* **64,** 894–899 (2007).
14. Lewis, C. M. *et al.* Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am. J. Hum. Genet.* **73,** 34–48 (2003).
15. Ruano, D. *et al.* Association of the gene encoding neurogranin with schizophrenia in males. *J. Psychiatr. Res.* **42,** 125–133 (2008).
16. Broadbelt, K., Ramprasaud, A. & Jones, L. B. Evidence of altered neurogranin immunoreactivity in areas 9 and 32 of schizophrenic prefrontal cortex. *Schizophr. Res.* **87,** 6–14 (2006).
17. Bernal, J., Rodriguez-Pena, A., Iniguez, M. A., Ibarrola, N. & Munoz, A. Influence of thyroid hormone on brain gene expression. *Acta Med. Austriaca* **19** (suppl. 1), 32–35 (1992).
18. Huang, F. L., Huang, K. P. & Boucheron, C. Long-term enrichment enhances the cognitive behavior of the aging neurogranin null mice without affecting their hippocampal LTP. *Learn. Mem.* **14,** 512–519 (2007).
19. Li, J., Pak, J. H., Huang, F. L. & Huang, K. P. *N*-methyl-D-aspartate induces neurogranin/RC3 oxidation in rat brain slices. *J. Biol. Chem.* **274,** 1294–1300 (1999).
20. Bliss, T. V. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361,** 31–39 (1993).
21. Toulopoulou, T. *et al.* Substantial genetic overlap between neurocognition and schizophrenia: genetic modeling in twin samples. *Arch. Gen. Psychiatry* **64,** 1348–1355 (2007).
22. Wang, H. *et al.* CaMKII activation state underlies synaptic labile phase of LTP and short-term memory formation. *Curr. Biol.* **18,** 1546–1554 (2008).
23. Flora, A., Garcia, J. J., Thaller, C. & Zoghbi, H. Y. The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proc. Natl Acad. Sci. USA* **104,** 15382–15387 (2007).
24. Pitt, D. & Hopkins, I. A syndrome of mental retardation, wide mouth and intermittent overbreathing. *Aust. Paediatr. J.* **14,** 182–184 (1978).
25. Kalscheuer, V. M. *et al.* Disruption of the *TCF4* gene in a girl with mental retardation but without the classical Pitt-Hopkins syndrome. *Am. J. Med. Genet. A.* **146A,** 2053–2059 (2008).
26. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genet.* **35,** 131–138 (2003).
27. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
28. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst.* **22,** 719–748 (1959).
29. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the subjects and their relatives and staff at the recruitment centres. This work was sponsored by EU grants LSHM-CT-2006-037761 (Project SGENE), PIAP-GA-2008-218251 (Project PsychGene) and HEALTH-F2-2009-223423 (Project PsychCNVs). Genotyping of the Dutch samples was sponsored by NIMH funding, R01 MH078075. This work was also supported by the National Genomic Network (NGFN-2) of the German Federal Ministry of Education and Research (BMBF) and Marie Curie grant PIAP-GA-2008-218251 (PsychGene). M.M.N. received support from the Alfried Krupp von Bohlen und Halbach-Stiftung. We are grateful to S. Schreiber and M. Krawczak for providing genotype data for PopGen controls, and to K.-H. Jöckel and R. Erbel for providing control individuals from the Heinz Nixdorf Recall Study. Recruitment of the patients from Munich was partially supported by GlaxoSmithKline. We are grateful to the Genetics Research Centre GmbH, an initiative by GlaxoSmithKline and LMU. The Northern Finland Birth Cohort 1966 (NFBC66) is thanked for providing population controls for the study. The genotyping of NFBC66 was financially supported by National Institutes of Health grant 1R01HL087679-01, STAMPEED.

Author Contributions H.S., S.S., D.A.C., D.S.C., D.R., E. Sigurdsson and K.S. wrote the first draft of the paper. M.H., B.B.M., P.M., I.G., H.-J.M., A.H., A.C.N., G.F., N.W., J.L., J. Suvisaari, A.T.-H., T.T., E.B., R.M., M.R., S. Tosato, S.D., I.M., J.O., O.A.A., M.R., R.A.O., L.A.K., O.G., A.D.B., M. Nyegaard, A.F.-J., M. Nordentoft, D.H., B.N.-P., Y.B., R.B., H.B.R., S. Timm, M.M., I.B., J.M.R., L.A., V.K., J. Sanjuan, R.F., E.V., U.E., M.P., J.L.Y., N.B.F., R.M.C., V.G., A.C., C.A., J.C., E.G.J., L.T., I.A., O.M., P.B.M., B.F., T.P. and GROUP recruited, diagnosed and gathered phenotypes. H.S., D.R., R.d.F.,

©2009 Macmillan Publishers Limited. All rights reserved

E. Strengman, T.S., P.M.M., T.T., J.R.G., U.T., H.P., D.B.G., T.W., D.A.C., L.P., A.K., D.S.C. and K.S. planned, supervised and coordinated the work. S.S., H.S., S.C., P.O., G.M., A.I., T.E.T., O.P.H.P., D.G., K.V.S., M.M.N., T.H. and A.K. analysed the data. All authors contributed to the current version of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.S. (kari.stefansson@decode.is).

Hreinn Stefansson[1]*, Roel A. Ophoff[2,3]*, Stacy Steinberg[1]*, Ole A. Andreassen[4], Sven Cichon[5], Dan Rujescu[6], Thomas Werge[7], Olli P. H. Pietiläinen[8,9], Ole Mors[10], Preben B. Mortensen[11], Engilbert Sigurdsson[12,13], Omar Gustafsson[1], Mette Nyegaard[14], Annamari Tuulio-Henriksson[15], Andres Ingason[1], Thomas Hansen[7], Jaana Suvisaari[15], Jouko Lonnqvist[15], Tiina Paunio[16], Anders D. Børglum[10,14], Annette Hartmann[6], Anders Fink-Jensen[17], Merete Nordentoft[18], David Hougaard[19], Bent Norgaard-Pedersen[19], Yvonne Böttcher[1], Jes Olesen[20], René Breuer[21], Hans-Jürgen Möller[22], Ina Giegling[6], Henrik B. Rasmussen[7], Sally Timm[23], Manuel Mattheisen[5], István Bitter[24], János M. Réthelyi[24], Brynja B. Magnusdottir[12,13], Thordur Sigmundsson[12,13], Pall Olason[1], Gisli, Masson[1], Jeffrey R. Gulcher[1], Magnus Haraldsson[12,13], Ragnheidur Fossdal[1], Thorgeir E. Thorgeirsson[1], Unnur Thorsteinsdottir[1,13], Mirella Ruggeri[25], Sarah Tosato[25], Barbara Franke[26], Eric Strengman[2], Lambertus A. Kiemeney[27], GROUP†, Ingrid Melle[4], Srdjan Djurovic[28], Lilia Abramova[29], Vasily Kaleda[29], Julio Sanjuan[30], Rosa de Frutos[31], Elvira Bramon[32], Evangelos Vassos[32,33], Gillian Fraser[34], Ulrich Ettinger[32,33], Marco Picchioni[32], Nicholas Walker[35], Timi Toulopoulou[33], Anna C. Need[36], Dongliang Ge[36], Joeng Lim Yoon[37], Kevin V. Shianna[36], Nelson B. Freimer[3], Rita M. Cantor[3,37], Robin Murray[32,33], Augustine Kong[1], Vera Golimbet[29], Angel Carracedo[38], Celso Arango[39], Javier Costas[40], Erik G. Jönsson[41], Lars Terenius[41], Ingrid Agartz[41], Hannes Petursson[12,13], Markus M. Nöthen[42], Marcella Rietschel[21], Paul M. Matthews[43], Pierandrea Muglia[44], Leena Peltonen[8,9], David St Clair[34], David B. Goldstein[36], Kari Stefansson[1,13] & David A. Collier[32,45]

[1]deCODE genetics, Sturlugata 8, IS-101 Reykjavik, Iceland. [2]Department of Medical Genetics and Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands. [3]UCLA Center for Neurobehavioral Genetics, Charles E. Young Drive South, Los Angeles, California 90024, USA. [4]Department of Psychiatry, Ullevål University Hospital and Institute of Psychiatry, University of Oslo, Kirkeveien 166, N-0407 Oslo, Norway. [5]Department of Genomics, Life and Brain Center, University of Bonn, Sigmund-Freud-Strasse 25, D-53127 Bonn, Germany. [6]Division of Molecular and Clinical Neurobiology, Department of Psychiatry, Ludwig-Maximilians-University, Nußbaumstrasse 7, 80336 Munich, Germany. [7]Research Institute of Biological Psychiatry, Mental Health Centre Sct. Hans Copenhagen University Hospital, DK-4000 Roskilde, Denmark. [8]Institute of Molecular Medicine, Biomedicum Helsinki, Haartmaninkatu 8, 00290 Helsinki, Finland. [9]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [10]Centre for Psychiatric Research, Aarhus University Hospital, Risskov, Skovagervej 2, 8240 Risskov, Denmark. [11]National Centre for Register-based Research, Aarhus University, Taasingegade 1, DK-8000 Aarhus, Denmark. [12]Department of Psychiatry, National University Hospital, Hringbraut, 101 Reykjavik, Iceland. [13]University of Iceland, School of Medicine, Laeknagardi, 101 Reykjavik, Iceland. [14]Department of Human Genetics, The Bartholin Building, Aarhus University, DK-8000 Arhus C, Denmark. [15]Department of Mental Health and Alcohol Research, National Public Health Institute, Mannerheimintie 166, FIN-00300 Helsinki, Finland. [16]Department for Molecular Medicine, National Public Health Institute, Biomedicum, Haartmaninkatu 8, 00290 Helsinki, Finland. [17]Mental Health Centre Rigshospitalet, Copenhagen University Hospital, DK-2100 Copenhagen Ø, Denmark. [18]Psychiatric Centre Bisbebjerg, Building 13A, Bispebjerg Hospital, Bispebjerg Bakke 23, 2400 Copenhagen NV, Denmark. [19]Section of Neonatal Screening and Hormones, Department Clinical Chemistry and Immunology, The State Serum Institute, Artillerivej 5, 2300 Copenhagen S, Denmark. [20]Department of Neurology, 57 Nordre Ringvej, Glostrup Hospital, Glostrup, DK-2600 Copenhagen, Denmark. [21]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, University of Heidelberg, J5, D-68159 Mannheim, Germany. [22]Department of Psychiatry, Ludwig-Maximilians-University, Nußbaumstrasse 7, 80336 Munich, Germany. [23]Mental Health Centre Frederiksberg, Copenhagen University Hospital, DK-2000 Frederiksberg, Denmark. [24]Semmelweis University, Department of Psychiatry and Psychotherapy, Budapest 1083, Hungary. [25]Section of Psychiatry and Clinical Psychology, University of Verona, Verona, 37134 Verona, Italy. [26]Department of Human Genetics, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. [27]Department of Epidemiology and Biostatistics and Department of Urology, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. [28]Department of Medical Genetics, Ulleval University Hospital and Institute of Psychiatry, University of Oslo, Kirkeveien 166, N-0407 Oslo, Norway. [29]Mental Health Research Center, Russian Academy of Medical Sciences, Zagorodnoe sh. 2/2, 117152 Moscow, Russia. [30]Unidad de Psiquiatría, Facultad de Medicina, Universidad de Valencia, CIBERSAM, 46010 Valencia, Spain. [31]Departamento de Genética. Facultad de Biología, Universidad de Valencia, CIBERSAM, Spain. [32]Division of Psychological Medicine, Institute of Psychiatry, King's College, London SE5 8AF, UK. [33]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, London SE5 8AF, UK. [34]Department of Mental Health, University of Aberdeen, Royal Cornhill Hospital, Aberdeen AB25 2ZD, UK. [35]Ravenscraig hospital, Inverkip Road, Greenock PA16 9HA, UK. [36]Institute for Genome Sciences and Policy, Center for Population Genomics and Pharmacogenetics, 4011 GSRB II 103 Research Drive, Duke University, DUMC Box 3471, Durham, North Carolina 27708, USA. [37]Department of Human Genetics, UCLA, 695 Charles Young Drive South, Los Angeles, California 90095, USA. [38]Fundación Pública Galega de Medicina Xenómica-Complexo Universitario Hospitalario de Santiago, and CIBER de Enfermedades Raras (CIBERER), IML-Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain. [39]Hospital General Universitario Gregorio Marañón, Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Madrid, Spain. [40]Fundación Pública Galega de Medicina Xenómica, and CIBER de Enfermedades Raras (CIBERER), 46010 Valencia, Spain. [41]Department of Clinical Neuroscience, HUBIN project, Karolinska Institutet and Hospital, R5:00, SE-171 76 Stockholm, Sweden. [42]Institute of Human Genetics, University of Bonn, Wilhelmstrasse 31, D-53111 Bonn, Germany. [43]Clinical Imaging Centre, Clinical Pharmacology and Discovery Medicine, GlaxoSmithKline, Hammersmith Hospital, London W12 ONN, UK. [44]Medical Genetics, GlaxoSmithKline R&D, Via A. Fleming 4, 37135 Verona, Italy. [45]Psychiatric Laboratory, Department of Psychiatry, West China Hospital, Sichuan University, 610065 Sichuan, China.
*These authors contributed equally to this work.

†Genetic Risk and Outcome in Psychosis (GROUP)

René S. Kahn[1], Don H. Linszen[2], Jim van Os[3], Durk Wiersma[4], Richard Bruggeman[4], Wiepke Cahn[1], Lieuwe de Haan[2], Lydia Krabbendam[3] & Inez Myin-Germeys[3]

[1]Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Postbus 85060, 3508 AB, Utrecht, The Netherlands. [2]Academic Medical Centre University of Amsterdam, Department of Psychiatry, Amsterdam, NL326 Groot-Amsterdam, The Netherlands. [3]Maastricht University Medical Centre, South Limburg Mental Health Research and Teaching Network, P. Debyelaan 25, 6229 HX Maastricht, Maastricht, The Netherlands. [4]University Medical Center Groningen, Department of Psychiatry, University of Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands.

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Schizophrenia is a severe mental disorder with a lifetime risk of about 1%, characterized by hallucinations, delusions and cognitive deficits, with heritability estimated at up to 80%[1,2]. We performed a genome-wide association study of 3,322 European individuals with schizophrenia and 3,587 controls. Here we show, using two analytic approaches, the extent to which common genetic variation underlies the risk of schizophrenia. First, we implicate the major histocompatibility complex. Second, we provide molecular genetic evidence for a substantial polygenic component to the risk of schizophrenia involving thousands of common alleles of very small effect. We show that this component also contributes to the risk of bipolar disorder, but not to several non-psychiatric diseases.

We genotyped the International Schizophrenia Consortium (ISC) case-control sample for up to ~1 million single nucleotide polymorphisms (SNPs), augmented by imputed common HapMap SNPs. In the genome-wide association study (GWAS; genomic control $\lambda_{GC} = 1.09$; Supplementary Table 1 and Supplementary Figs 1–3), the most associated genotyped SNP ($P = 3.4 \times 10^{-7}$) was located in the first intron of myosin XVIIIB (*MYO18B*) on chromosome 22. The second strongest association comprised more than 450 SNPs on chromosome 6p spanning the major histocompatibility complex (MHC; Fig. 1). There is some evidence for between-site heterogeneity in both allele frequencies and odds ratios (Table 1). We observed associations consistent with previous reports in the 22q11.2 deletion region and *ZNF804A* (ref. 3) (Supplementary Table 2, Supplementary Fig. 2 and section 5 and 6 in Supplementary Information).

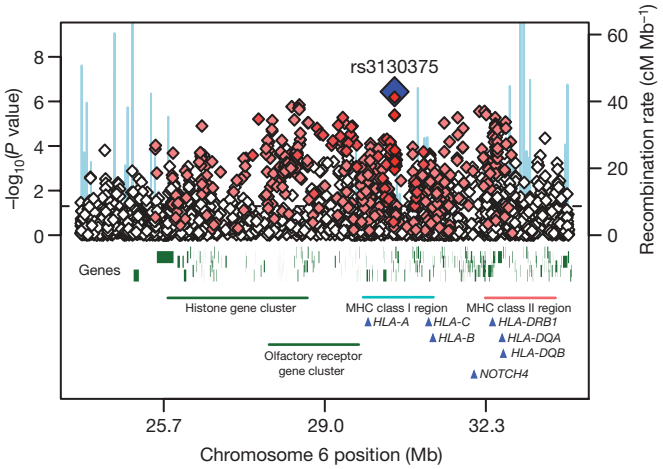


Figure 1 | Association results across the MHC region. Results are shown as $-\log_{10}$(*P* value) for genotyped SNPs. The most associated SNP is shown as a blue diamond. The colour of the remaining markers reflects r^2 with rs3130375, light pink, $r^2 > 0.1$, red, $r^2 > 0.8$. The recombination rate from the CEU HapMap (second *y* axis) is plotted in light blue.

The best imputed SNP, which reached genome-wide significance (rs3130297, $P = 4.79 \times 10^{-8}$, T allele odds ratio = 0.747, minor allele frequency (MAF) = 0.114, 32.3 megabases (Mb)), was also in the MHC, 7 kilobases (kb) from *NOTCH4*, a gene with previously reported associations with schizophrenia[4]. We imputed classical human leukocyte antigen (HLA) alleles; six were significant at $P < 10^{-3}$, found on the ancestral European haplotype[5] (Table 1, Supplementary Table 3 and section 3 in Supplementary Information). However, it was not possible to ascribe the association to a specific HLA allele, haplotype or region (Supplementary Table 3 and Supplementary Fig. 4).

We exchanged GWAS summary results with the Molecular Genetics of Schizophrenia (MGS) and SGENE consortia for genotyped SNPs with $P < 10^{-3}$. There were 8,008 cases and 19,077 controls of European descent in the combined sample (see refs 6, 7 and section 7 in Supplementary Information). Our top genotyped MHC SNP (rs3130375) had $P = 0.086$ and $P = 0.14$ in MGS and SGENE, respectively. Considering the combined results for genotyped and imputed SNPs across the MHC region more broadly, rs13194053 had a genome-wide significant combined $P = 9.5 \times 10^{-9}$ (ISC, MGS and SGENE: $P = 3 \times 10^{-4}$, 1×10^{-2} and 1×10^{-4}, respectively; C allele

Table 1 | MHC association for the most significant genotyped SNP rs3130375

a MHC association for rs3130375 by sample

Sample	Ancestry	Frequency (rs3130375, A allele) Cases	Controls	*P* value
University of Aberdeen	Scottish	0.132	0.168	0.0060
University of Edinburgh	Scottish	0.137	0.135	0.8930
University College London*	British	0.132	0.143	0.4836
Trinity College Dublin	Irish	0.110	0.170	0.0012
Cardiff University	Bulgarian	0.077	0.084	0.5602
Portuguese Island Collection	Portuguese	0.048	0.061	0.3510
Karolinska Institutet (5.0)	Swedish	0.043	0.119	0.0004
Karolinska Institutet (6.0)	Swedish	0.089	0.142	0.0040

b MHC association for classical HLA alleles with $P < 1 \times 10^{-3}$

HLA allele	Frequency†	Odds ratio	*P* value
*HLA-A*0101*	0.103	0.785	4×10^{-5}
*HLA-C*0701*	0.113	0.778	5×10^{-5}
*HLA-B*0801*	0.068	0.757	3×10^{-5}
*HLA-DRB*0301*	0.121	0.768	3×10^{-6}
*HLA-DQB*0201*	0.210	0.857	4×10^{-4}
*HLA-DQA*0501*	0.205	0.798	6×10^{-7}

Total sample Cochran–Mantel–Haenszel $P = 4 \times 10^{-7}$; Breslow–Day heterogeneity test $P = 0.012$ (d.f. = 6).
* SNP failed genotyping quality control in UCL. Allele frequency for UCL based on imputed genotypes.
† Frequency is estimated population frequency.

*Lists of authors and their affiliations appear at the end of the paper.

©2009 Macmillan Publishers Limited. All rights reserved

odds ratio = 0.82, 0.88 and 0.78) and was in linkage disequilibrium with rs3130375 ($r^2 = 0.35$ in HapMap). Across the region, 11 other SNPs had $P < 10^{-7}$ at 27.1–27.3 Mb and 32.7 Mb (Supplementary Table 5).

Our second approach was to evaluate whether common variants have an important role en masse, directly testing the classic theory of polygenic inheritance[8], previously hypothesized to apply to schizophrenia[9]. Although our GWAS analysis did not identify a large number of strongly associated loci, there could still be potentially thousands of very small individual effects that collectively account for a substantial proportion of variation in risk. We summarized variation across nominally associated loci into quantitative scores, and related the scores to disease state in independent samples[10]. Although variants of small effect (for example, genotypic relative risk (GRR) = 1.05) are unlikely to achieve even nominally significant P values, increasing proportions will be detected at increasingly liberal significance thresholds (P_T), for example, $P_T < 0.1$ or $P_T < 0.5$. Using such thresholds, we defined large sets of 'score alleles' in a discovery sample, to generate aggregate risk scores for individuals in independent target samples. We use the term score, instead of risk, as we cannot differentiate the minority of true risk alleles from unassociated variants.

We performed the score analyses on a reduced set of SNPs to facilitate analysis and interpretation. After filtering on MAF, genotyping rate and linkage disequilibrium (independent of association with schizophrenia), we obtained a subset of 74,062 autosomal SNPs in approximate linkage equilibrium (Supplementary Tables 6 and 7). In each discovery sample, we selected sets of score alleles at different association test P_T thresholds. For each individual in the target sample, we calculated the number of score alleles they possessed, each weighted by the log odds ratio from the discovery sample. To assess whether the aggregate scores reflect schizophrenia risk, we tested for a higher mean score in target cases compared to controls (sections 9–11 in Supplementary Information and Supplementary Table 7).

We selected males (2,176 cases, 1,642 controls) and females (1,146 cases, 1,945 controls) to form arbitrary discovery and target samples (Supplementary Table 8). Score alleles designated in the discovery sample were significantly enriched among target cases, and the effect was larger for increasingly liberal P_T thresholds. The score on the basis of all SNPs with male discovery $P_T < 0.5$ ($n = 37{,}655$ SNPs) was highly correlated with schizophrenia in target females ($P = 9 \times 10^{-19}$), explaining ~3% of the variance (Nagelkerke's pseudo R^2 from logistic regression), with higher scores in cases. The results were not driven by only a few highly associated regions (section 12 in Supplementary Information).

We eliminated several possible confounders, with emphasis on subtle population stratification (Supplementary Tables 9–15). Defining score alleles in British Isles samples and testing in target samples from Sweden, Portugal and Bulgaria, and vice versa, we observed a similar pattern of results. It is unlikely that the same substructure is overrepresented in the corresponding phenotype class when discovery and target samples are from distinct populations. The effect is also stronger for SNPs within annotated genes (Supplementary Table 16).

We used independent GWAS samples to replicate the polygenic component, to examine whether this component is shared with bipolar disorder[11], and to demonstrate specificity by considering non-psychiatric diseases. We used the entire ISC for the discovery sample, considering the five most informative P_T thresholds from the intra-ISC analyses. The independent target samples were the MGS European-American (MGS-EA), the MGS African-American (MGS-AA) and the UK sample described previously by O'Donovan *et al*[3]. The ISC-derived score was highly associated with disease in both European schizophrenia samples (Fig. 2, Supplementary Fig. 6 and Supplementary Table 17). The MGS-EA had a significantly higher mean $P_T < 0.5$ score in cases compared to controls ($P = 2 \times 10^{-28}$, $R^2 = 3.2\%$), as did the smaller O'Donovan sample ($P = 5 \times 10^{-11}$,

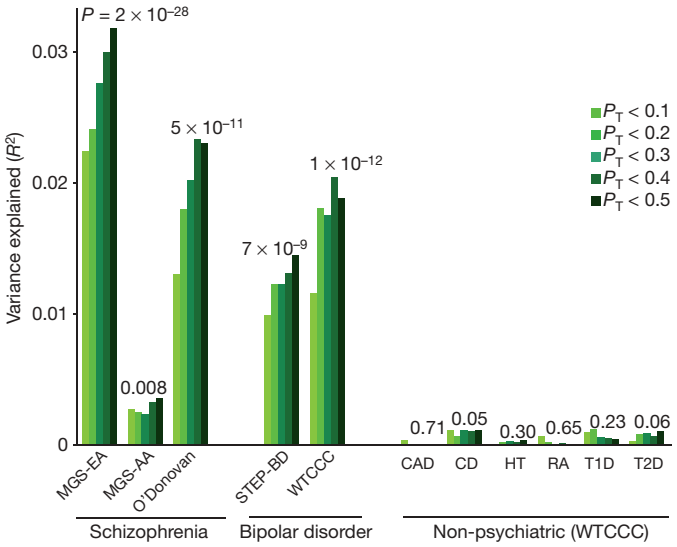


Figure 2 | Replication of the ISC-derived polygenic component in independent schizophrenia and bipolar disorder samples. Variance explained in the target samples on the basis of scores derived in the entire ISC for five significance thresholds ($P_T < 0.1$, 0.2, 0.3, 0.4 and 0.5, plotted left to right in each study). The y axis indicates Nagelkerke's pseudo R^2; the number above each set of bars is the P value for the $P_T < 0.5$ target sample analysis. CAD, coronary artery disease; CD, Crohn's disease; HT, hypertension; RA, rheumatoid arthritis; T1D, type I diabetes; T2D, type II diabetes. Numbers for cases/controls: MGS-EA 2,687/2,656; MGS-AA 1,287/973; O'Donovan 479/2,938; STEP-BD 955/1,498; WTCCC 1,829/2,935; CAD 1,926/2,935; CD 1,748/2,935; HT 1,952/2,935; RA 1,860/2,935; T1D 1,963/2,935; and T2D 1,924/2,935.

$R^2 = 2.3\%$). Aggregate differences in allele frequencies and patterns of linkage disequilibrium between Europeans and African-Americans are expected to lead to an attenuated effect. Still, MGS-AA cases carried more of the European-derived score alleles than the MGS-AA controls ($P = 0.008$; $R^2 = 0.4\%$).

The ISC-derived score alleles were also associated with bipolar disorder in two independent samples. Both samples, STEP-BD[12] and WTCCC[13], had higher mean $P_T < 0.5$ scores in cases than in controls ($P = 7 \times 10^{-9}$, $R^2 = 1.9\%$, and $P = 1 \times 10^{-12}$, $R^2 = 1.4\%$, respectively) indicating a substantial, shared genetic component.

To test disease specificity, we selected all six non-psychiatric WTCCC samples (coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I and type II diabetes). Controls are shared among the WTCCC case samples, including bipolar disorder. In contrast to schizophrenia and bipolar disorder, there was no association ($P > 0.05$) between the ISC-derived schizophrenia scores and these non-psychiatric diseases, for any P_T threshold.

We next investigated the genetic models consistent with our data. The total additive genetic variance (V_A) reflects the number of causal alleles, as well as their frequency and effect size distributions. However, the variance explained by the markers that tag these causal alleles (V_M) will be attenuated, reflecting the average extent of linkage disequilibrium between marker and causal allele. In our target samples, the variance explained by the observed score alleles (V_S) will be further attenuated by sampling variation and P_T threshold, such that $V_S \leq V_M \leq V_A$.

We used simulation to estimate possible values for V_M and V_A, by identifying models that produced profiles of V_S across P_T threshold that were similar to those observed in the ISC data, as indexed by the target sample R^2. Under a variety of genetic models, we simulated discovery and target data sets of comparable sample size to the ISC. On the basis of the empirical allele frequency distribution, we simulated marker SNPs, varying the proportion that were in linkage disequilibrium with causal variants, for which we varied allele frequency (uniform, U-shaped) and effect size distributions (fixed

©2009 Macmillan Publishers Limited. All rights reserved

GRR values, exponential GRR values, or fixed variance explained) as well as the extent of linkage disequilibrium (section 16 in Supplementary Information).

From a broad range of models, a subset produced results consistent with the ISC data (Fig. 3 and Supplementary Fig. 7). Among these, all led to similar estimates of V_M (mean 34%, range 32% to 36%). In models in which the causal alleles were imperfectly tagged ($r^2 < 1$), estimates of V_A can be considerably larger. Therefore, our estimate that common polygenic variation accounts for one-third of the total variation in schizophrenia risk is a lower bound for the true value, which could be much higher. Figure 3b shows seven examples from the range of consistent models, detailed in Supplementary Table 18.

The simulated models consistent with our observed results all indicated a substantial number of common variants, whereas models that invoked only a few common variants of large effect or only rare variants were not able to account for our findings. For example, if $V_M \approx 34\%$ arose from only 100 common causal alleles, with GRR values at the tagging marker between ~1.2–1.5, most would be detected at $P_T < 0.01$, and so the variance explained would decline, not increase, as more SNPs were added (Fig. 3c and Supplementary Table 19). It is possible that an observed GRR of ~1.05 could represent a large effect of a weakly tagged rare variant, for example, a tenfold effect of a 1/10,000 variant in complete linkage disequilibrium ($D' = 1$, but low r^2) with a genotyped SNP. However, as this would only hold for low frequency markers (MAF < ~0.1), we stratified our analysis by score allele frequency (Fig. 4a). For simulated models in which all causal variants were of low frequency (<0.05), a stratified analysis revealed the expected, skewed distribution (Fig. 4c and section 17 in Supplementary Information), which was more pronounced for rarer causal alleles, for example, 1/1,000 (data not shown). In contrast, models in which causal alleles followed a uniform frequency distribution provided a closer fit to our data (Fig. 4b; although note some enrichment in the second quintile, of ~13–35% score alleles). Moreover, rare variants are likely to be population specific and if recurrent, in linkage disequilibrium with different common alleles within and between populations. As such, they could not account for the observation of disease variation that is largely shared across our different populations.

Decreased reproductive fitness in schizophrenia[14] suggests that risk alleles of large to moderate effect will be under negative selection and therefore very rare[15,16]. This is not inconsistent with our results, because

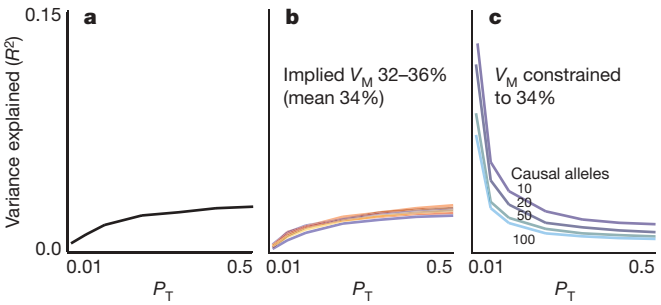


Figure 3 | Observed and simulated profiles of target sample variance explained. **a**, The observed variance explained is shown (R^2, black line). **b**, A subset of models that produced results consistent with the observed data is shown. All yielded similar estimates of the total variance explained by the SNPs that tag the causal variants, V_M, with a mean value of 34%. The seven models (shown as percentage SNPs, mean GRR/variance explained (V) per causal allele, linkage disequilibrium, and frequency model) were: M_1: 6.25%, GRR = 1.05, $r^2 = 1$, empirical; M_2: 25%, GRR = 1.025, $r^2 = 1$, empirical; M_3: 12%, GRR = 1.05, $r^2 < 1$, uniform; M_4: 32%, GRR = 1.04, $r^2 < 1$, U-shaped; M_5: 11%, $V = 0.00006$, $r^2 = 1$, empirical; M_6: 25%, GRR(exponential) = 1.025, $r^2 < 1$, uniform; M_7: 100%, GRR(exponential) = 1.012, $r^2 < 1$, uniform. **c**, Four inconsistent models with fewer variants of larger effect are shown.

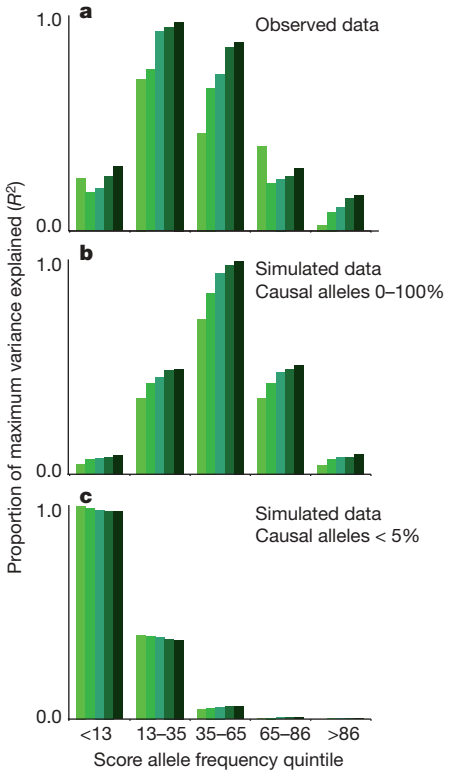


Figure 4 | Analysis stratified by score allele frequency. **a**, The observed data for the ISC/MGS-EA comparison is shown. The *y* axis is the target sample pseudo R^2, scaled within each figure as a proportion of the maximum value observed for five significance thresholds ($P_T < 0.1$, 0.2, 0.3, 0.4 and 0.5, plotted left to right in each quintile). **b**, **c**, Shown are results for simulated data: the common variant model, with a uniform frequency distribution for causal risk-increasing alleles (**b**) and a multiple rare variant model, in which the collective frequency of rare variants at a locus that all reside on the same haplotypic background with respect to the genotyped SNP was bounded at a maximum of 5% (**c**).

the common variants indexed by our polygenic score will not be subjected to strong selection, by virtue of their very small individual effect sizes. Our results do not exclude important contributions of rare variants for schizophrenia[15], because rare variants are expected as part of the allele frequency/effect size spectrum of a polygenic model. We and others recently reported higher genome-wide rates of rare copy number variants in schizophrenia[17–19]. However, our results indicate that medical sequencing and studies of structural variation to identify rare, highly penetrant variants will not alone fully characterize the genetic risk factors.

In conclusion, our molecular genetic data strongly support a polygenic basis to schizophrenia that (1) involves common SNPs, (2) explains at least one-third of the total variation in liability, (3) is substantially shared with bipolar disorder, and (4) is largely not shared with several non-psychiatric diseases. We also identified variants in the MHC region that received support in two independent studies, although the population specificity and extensive linkage disequilibrium will make follow-up challenging.

A highly polygenic model suggests that genetically influenced individual differences across domains of brain development and function may form a diathesis for major psychiatric illness, perhaps as multiple growth and metabolic pathways influence human height[20]. Our results may also reflect heterogeneity, such that some patients have aetiologically distinct diseases. The shared genetic liability between schizophrenia and bipolar disorder, previously suggested by clinical and genetic epidemiology[11,21], opens up the possibility of genetically based refinements in diagnosis. However, the scores derived here have little value for individual risk prediction, meaning that application

©2009 Macmillan Publishers Limited. All rights reserved

to clinical genetic testing for schizophrenia would be unwarranted. In the future, measures of polygenic burden, along with known risk loci and non-genetic factors such as season of birth, life stress, obstetrical complications, viral infections and epigenetics, could open new avenues for studying gene–gene and gene–environment interactions.

Increasing the discovery sample size should substantially refine the polygenic scores derived here. The variance explained by the observed score increases from ~3% to over 20% in extended simulations of 20,000 case/control pairs, as will soon be available by international meta-analytic efforts such as the Psychiatric GWAS Consortium[22–24] (section 18 in Supplementary Information and Supplementary Fig. 8). Furthermore, analyses that focus on gene pathways, clinical features and non-additivity may increase the variance captured by the score and identify genes or biological systems that are either shared by, or unique to, schizophrenia and bipolar disorder.

We identified fewer unambiguously associated variants than studies of some non-psychiatric diseases of comparable size[25]. Nonetheless, for other diseases replicated variants typically account for only a modest fraction of risk. The nature of this 'missing heritability' is a general problem now faced by complex disease geneticists[26]. For schizophrenia, our data point to a genetic architecture that includes many common variants of small effect. The extent to which similar models characterize genetic variation within and across other complex diseases remains to be investigated.

METHODS SUMMARY

Cases satisfied criteria for schizophrenia. Clinical characteristics and copy number variation have been described previously[17]. DNA was extracted from whole blood, with approval from institutional review boards. Genotypes were called using the Birdseed/Birdsuite algorithm[27] and analyses were performed with PLINK version 1.05 (ref. 28). Association analyses used a Cochran–Mantel–Haenszel test and logistic regression with covariates for sample site and ancestry. In the simulations, we generated data sets with pairs of unobserved variants and marker SNPs in varying degrees of within-pair linkage disequilibrium, on the basis of the effective number of independent SNPs in the ISC, and assuming Hardy–Weinberg equilibrium and linkage equilibrium between different pairs of SNPs. We considered a large grid of possible values for allele frequency and effect size distributions, also varying the proportion of non-null variants and the linkage disequilibrium between causal allele and observed marker. We retained models that produced similar profiles of target sample R^2 compared to the original ISC analysis, for the same range of P_T thresholds, and calculated the indicated total genetic variance under these models, assuming additivity within and across loci. See Supplementary Information for details.

Received 11 February; accepted 8 June 2009.
Published online 1 July; corrected 6 August 2009 (see full-text HTML version for details).

1. Cardno, A. G. & Gottesman, I. I. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am. J. Med. Genet.* **97,** 12–17 (2000).
2. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60,** 1187–1192 (2003).
3. O'Donovan, M. C. *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genet.* **40,** 1053–1055 (2008).
4. Wei, J. & Hemmings, G. P. The *NOTCH4* locus is associated with susceptibility to schizophrenia. *Nature Genet.* **25,** 376–377 (2000).
5. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60,** 1–18 (2008).
6. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* doi:10.1038/nature08186 (this issue).
7. Douglas, J. S. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* doi:10.1038/nature08192 (this issue).
8. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Philos. Trans. R. Soc. Edinb.* **52,** 399–433 (1918).
9. Gottesman, I. I. & Shields, J. A polygenic theory of schizophrenia. *Proc. Natl Acad. Sci. USA* **58,** 199–205 (1967).
10. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17,** 1520–1528 (2007).
11. Craddock, N., O'Donovan, M. C. & Owen, M. J. Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. *Schizophr. Bull.* **32,** 9–16 (2006).
12. Sklar, P. *et al.* Whole-genome association study of bipolar disorder. *Mol. Psychiatry* **13,** 558–569 (2008).
13. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
14. Svensson, A. C., Lichtenstein, P., Sandin, S. & Hultman, C. M. Fertility of first-degree relatives of patients with schizophrenia: a three generation perspective. *Schizophr. Res.* **91,** 238–245 (2007).
15. McClellan, J. M., Susser, E. & King, M. C. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190,** 194–199 (2007).
16. Craddock, N., O'Donovan, M. C. & Owen, M. J. Phenotypic and genetic complexity of psychosis. Invited commentary on. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190,** 200–203 (2007).
17. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008).
18. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).
19. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nature Genet.* **40,** 880–885 (2008).
20. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40,** 575–583 (2008).
21. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373,** 234–239 (2009).
22. Psychiatric GWAS Consortium Steering Committee. A framework for interpreting genome-wide association studies of psychiatric disorders. *Mol. Psychiatry* **14,** 10–17 (2009).
23. Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: history, rationale and prospects for psychiatric disorders. *Am. J. Psychiatry* **166,** 540–556 (2009).
24. Cross Disorder Phenotype Group of the Psychiatric GWAS Consortium. Dissecting the phenotype in genome-wide association studies of psychiatric illness. *Br. J. Psychiatry* doi:10.1192/bjp.bp.108.063156 (in the press).
25. Manolio, T. A., Brooks, L. D. & Collins, F. S. A. HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
26. Maher, B. The case of the missing heritability. *Nature* **456,** 18–21 (2008).
27. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genet.* **40,** 1253–1260 (2008).
28. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the patients and families who contributed to these studies. We also thank E. Lander, N. Patterson and members of the Medical and Population Genetics group at the Broad Institute of Harvard and Massachusetts Institute of Technology for valuable discussion, and members of the Broad Biological Samples and Genetic Analysis Platforms for sample management and genotyping. We particularly thank D. Levinson and P. Gejman for allowing access to the MGS samples, and J. Shi for analytic support with the MGS samples. The group at the Stanley Center for Psychiatric Research at the Broad Institute was supported by the Stanley Medical Research Institute (E.M.S.), the Sylvan C. Herman Foundation (E.M.S.), and MH071681 (P.S.). The Cardiff University group was supported by a Medical Research Council (UK) Programme grant and the National Institutes of Mental Health (USA) (CONTE: 2 P50 MH066392-05A1). The group at the Karolinska Institutet was supported by the Swedish Council for Working Life and Social Research (FO 184/2000; 2001-2368). The Massachusetts General Hospital group was supported by the Stanley Medical Research Institute (P.S.), MH071681 and MH077139 (P.S.) and a Narsad Young Investigator Award (S.M.P.). The group at the Queensland Institute of Medical Research was supported by the Australian National Health and Medical Research Council (grants 389892, 442915, 496688 and 496674) and thanks S. Gordon for data preparation. The Trinity College Dublin group was supported by Science Foundation Ireland, the Health Research Board (Ireland), the Stanley Medical Research Institute and the Wellcome Trust; Irish controls were supplied by J. McPartlin from the Trinity College Biobank. The work at the University of Aberdeen was partly funded by GlaxoSmithKline and Generation Scotland, Genetics Health Initiative. University College London clinical and control samples were collected with support from the Neuroscience Research Charitable Trust, the Camden and Islington Mental Health and Social Care Trust, East London and City Mental Heath Trust, the West Berkshire NHS Trust, the West London Mental Health Trust, Oxfordshire and Buckinghamshire Mental Health Partnership NHS Trust, South Essex Partnership NHS Foundation Trust, Gloucestershire Partnership NHS Foundation Trust, Mersey Care NHS Trust, Hampshire Partnership NHS Trust and the North East London Mental Health Trust. The collection of the University of Edinburgh cohort was supported by the Wellcome Trust Clinical Research Facility (Edinburgh) and grants from The Wellcome Trust, London and the Chief Scientist Office of the Scottish Government. The group at the University of North Carolina, Chapel Hill, was supported by MH074027, MH077139 and MH080403, the Sylvan C. Herman Foundation (P.F.S.) and the Stanley Medical Research Institute (P.F.S.). The group at the University of Southern California thanks the patients and their families for their collaboration, and acknowledges the support of the National Institutes of Mental Health and the Department of Veterans Affairs.

©2009 Macmillan Publishers Limited. All rights reserved

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.S. (sklar@chgr.mgh.harvard.edu) and S.M.P. (shaun@pngu.mgh.harvard.edu).

The International Schizophrenia Consortium

Manuscript preparation Shaun M. Purcell[1,2,3,4], Naomi R. Wray[5], Jennifer L. Stone[1,2,3,4], Peter M. Visscher[5], Michael C. O'Donovan[6], Patrick F. Sullivan[7], Pamela Sklar[1,2,3,4]; **Data analysis** Shaun M. Purcell[1,2,3,4] (Leader), Jennifer L. Stone[1,2,3,4], GWAS analysis subgroup: Patrick F. Sullivan[7], Douglas M. Ruderfer[1,2,3,4], Andrew McQuillin[8], Derek W. Morris[9], Colm T. O'Dushlaine[9], Aiden Corvin[9], Peter A. Holmans[6], Michael C. O'Donovan[6], Pamela Sklar[1,2,3,4], Polygene analyses subgroup: Naomi R. Wray[5], Stuart Macgregor[5], Pamela Sklar[1,2,3,4], Patrick F. Sullivan[7], Michael C. O'Donovan[6], Peter M. Visscher[5]; **Management committee** Hugh Gurling[8], Douglas H. R. Blackwood[10], Aiden Corvin[9], Nick J. Craddock[6], Michael Gill[9], Christina M. Hultman[11,12], George K. Kirov[6], Paul Lichtenstein[11], Andrew McQuillin[8], Walter J. Muir[10], Michael C. O'Donovan[6], Michael J. Owen[6], Carlos N. Pato[13], Shaun M. Purcell[1,2,3,4], Edward M. Scolnick[2,3], David St Clair[14], Jennifer L. Stone[1,2,3,4], Patrick F. Sullivan[7], Pamela Sklar[1,2,3,4] (Leader); **Cardiff University** Michael C. O'Donovan[6], George K. Kirov[6], Nick J. Craddock[6], Peter A. Holmans[6], Nigel M. Williams[6], Lyudmila Georgieva[6], Ivan Nikolov[6], N. Norton[6], H. Williams[6], Draga Toncheva[16], Vihra Milanova[17], Michael J. Owen[6]; **Karolinska Institutet/University of North Carolina at Chapel Hill** Christina M. Hultman[11,12], Paul Lichtenstein[11], Emma F. Thelander[11], Patrick Sullivan[7]; **Trinity College Dublin** Derek W. Morris[9], Colm T. O'Dushlaine[9], Elaine Kenny[9], Emma M. Quinn[9], Michael Gill[9], Aiden Corvin[9]; **University College London** Andrew McQuillin[8], Khalid Choudhury[8], Susmita Datta[8], Jonathan Pimm[8], Srinivasa Thirumalai[18], Vinay Puri[8], Robert Krasucki[8], Jacob Lawrence[8], Digby Quested[19], Nicholas Bass[8], Hugh Gurling[8]; **University of Aberdeen** Caroline Crombie[15], Gillian Fraser[15], Soh Leh Kuan[14], Nicholas Walker[20], David St Clair[14]; **University of Edinburgh** Douglas H. R. Blackwood[10], Walter J. Muir[10], Kevin A. McGhee[10], Ben Pickard[10], Pat Malloy[10], Alan W. Maclean[10], Margaret Van Beck[10]; **Queensland Institute of Medical Research** Naomi R. Wray[5], Stuart Macgregor[5], Peter M. Visscher[5]; **University of Southern California** Michele T. Pato[13], Helena Medeiros[13], Frank Middleton[21], Celia Carvalho[13], Christopher Morley[21], Ayman Fanous[13,22,23,24], David Conti[13], James A. Knowles[13], Carlos Paz Ferreira[25], Antonio Macedo[26], M. Helena Azevedo[26], Carlos N. Pato[13]; **Massachusetts General Hospital** Jennifer L. Stone[1,2,3,4], Douglas M. Ruderfer[1,2,3,4], Andrew N. Kirby[2,3,4], Manuel A. R. Ferreira[1,2,3,4], Mark J. Daly[2,3,4], Shaun M. Purcell[1,2,3,4], Pamela Sklar[1,2,3,4]; **Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard** Shaun M. Purcell[1,2,3,4], Jennifer L. Stone[1,2,3,4], Kimberly Chambert[3,4], Douglas M. Ruderfer[1,2,3,4], Finny Kuruvilla[4], Stacey B. Gabriel[4], Kristin Ardlie[4], Jennifer L. Moran[4], Mark J. Daly[2,3,4], Edward M. Scolnick[3,4], Pamela Sklar[1,2,3,4]

[1]Psychiatric and Neurodevelopmental Genetics Unit, [2]Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [3]Stanley Center for Psychiatric Research, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [4]The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [5]Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia. [6]MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine, School of Medicine, Cardiff University, Cardiff C14 4XN, UK. [7]Departments of Genetics, Psychiatry, and Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [8]Molecular Psychiatry Laboratory, Research Department of Mental Health Sciences, University College London Medical School, Windeyer Institute of Medical Sciences, 46 Cleveland Street, London W1T 4JF, UK. [9]Neuropsychiatric Genetics Research Group, Department of Psychiatry and Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Ireland. [10]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK. [11]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [12]Department of Neuroscience, Psychiatry, Ulleråker, Uppsala University, SE-750 17 Uppsala, Sweden. [13]Center for Genomic Psychiatry, University of Southern California, Los Angeles, California 90033, USA. [14]Institute of Medical Sciences, [15]Department of Mental Health, University of Aberdeen, Aberdeen AB25 2ZD, UK.[16]Department of Medical Genetics, University Hospital Maichin Dom, Sofia 1431, Bulgaria. [17]Department of Psychiatry, First Psychiatric Clinic, Alexander University Hospital, Sofia 1431, Bulgaria. [18]West Berkshire NHS Trust, 25 Erleigh Road, Reading RG3 5LR, UK. [19]Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford OX3 7JX, UK. [20]Ravenscraig Hospital, Inverkip Road, Greenock PA16 9HA, UK. [21]State University of New York – Upstate Medical University, Syracuse, New York 13210, USA. [22]Washington VA Medical Center, Washington DC 20422, USA. [23]Department of Psychiatry, Georgetown University School of Medicine, Washington DC 20057, USA. [24]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia 23298, USA. [25]Department of Psychiatry, Sao Miguel, 9500-310 Azores, Portugal. [26]Department of Psychiatry University of Coimbra, 3004-504 Coimbra, Portugal.

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

Common variants on chromosome 6p22.1 are associated with schizophrenia

Jianxin Shi[1], Douglas F. Levinson[1], Jubao Duan[2], Alan R. Sanders[2], Yonglan Zheng[2], Itsik Pe'er[3], Frank Dudbridge[4], Peter A. Holmans[5], Alice S. Whittemore[6], Bryan J. Mowry[7], Ann Olincy[8], Farooq Amin[9], C. Robert Cloninger[10], Jeremy M. Silverman[11], Nancy G. Buccola[12], William F. Byerley[13], Donald W. Black[14], Raymond R. Crowe[14], Jorge R. Oksenberg[15], Daniel B. Mirel[16], Kenneth S. Kendler[17], Robert Freedman[8] & Pablo V. Gejman[2]

Schizophrenia, a devastating psychiatric disorder, has a prevalence of 0.5–1%, with high heritability (80–85%) and complex transmission[1]. Recent studies implicate rare, large, high-penetrance copy number variants in some cases[2], but the genes or biological mechanisms that underlie susceptibility are not known. Here we show that schizophrenia is significantly associated with single nucleotide polymorphisms (SNPs) in the extended major histocompatibility complex region on chromosome 6. We carried out a genome-wide association study of common SNPs in the Molecular Genetics of Schizophrenia (MGS) case-control sample, and then a meta-analysis of data from the MGS, International Schizophrenia Consortium and SGENE data sets. No MGS finding achieved genome-wide statistical significance. In the meta-analysis of European-ancestry subjects (8,008 cases, 19,077 controls), significant association with schizophrenia was observed in a region of linkage disequilibrium on chromosome 6p22.1 ($P = 9.54 \times 10^{-9}$). This region includes a histone gene cluster and several immunity-related genes—possibly implicating aetiological mechanisms involving chromatin modification, transcriptional regulation, autoimmunity and/or infection. These results demonstrate that common schizophrenia susceptibility alleles can be detected. The characterization of these signals will suggest important directions for research on susceptibility mechanisms.

The symptoms and course of schizophrenia are variable, without forming distinct familial subtypes[1]. There are positive (delusions and hallucinations), negative (reduced emotions, speech and interest), and disorganized (disrupted syntax and behaviour) symptoms, as well as mood symptoms in many cases. Onset is typically in adolescence or early adulthood, and rarely in childhood. The course of illness can range from acute episodes with primarily positive symptoms to the more common chronic or relapsing patterns often accompanied by cognitive disability and histories of childhood conduct or developmental disorders.

To search for common schizophrenia susceptibility variants, we carried out a genome-wide association study (GWAS) in cases from three methodologically similar National Institute of Mental Health repository-based studies, and screened controls from the general population. Cases were included with diagnoses of schizophrenia or (in 10% of cases) schizoaffective disorder, with the schizophrenia syndrome present for at least six months, genotyped with the Affymetrix 6.0 array. Because the frequencies of tag SNPs and disease susceptibility alleles can vary across populations, we carried out a primary analysis of the larger MGS European-ancestry sample (2,681 cases, 2,653 controls), and then further analyses of the African-American sample (1,286 cases, 973 controls) and of both of these samples combined, to test the hypothesis that there are alleles that influence susceptibility in both populations. All association tests were corrected using principal component scores indexing subjects' ancestral origins. Genotypic data were imputed for additional HapMap SNPs in selected regions.

These analyses did not produce genome-wide significant findings at a threshold of $P < 5 \times 10^{-8}$ (see Supplementary Methods). Table 1 summarizes the best results in the European-ancestry and African-American analyses. The strongest genic findings were in *CENTG2* (also known as AGAP1; chromosome 2q37.2, $P = 4.59 \times 10^{-7}$) in European-ancestry subjects, and in *ERBB4* (2q34, $P = 2.14 \times 10^{-6}$) in African-American subjects. Common variants in *ERBB4* (the strongest signal in African-American subjects) and its ligand neuregulin 1 (*NRG1*) have been reported to be associated with schizophrenia[3]. Further information about results in previously reported schizophrenia candidate genes is provided in Supplementary Results 3 and Supplementary Data 2.

As shown in Supplementary Table 17, power was adequate in the MGS European-ancestry sample to detect very common risk alleles (30–60% frequency, log additive effects) with genotypic relative risks of approximately 1.3, with lower power in the smaller African-American sample. The results indicate that there are few or no single common loci with such large effects on risk. The lack of consistency between the European-ancestry and African-American analyses could be due to low power, but new genome-wide analyses presented in the companion paper by the International Schizophrenia Consortium (ISC)[4] (discussed further later) suggest that although there is substantial overlap between the sets of risk alleles that are detected by GWAS in pairs of European-ancestry samples, much less

[1]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California 94304, USA. [2]Center for Psychiatric Genetics, NorthShore University HealthSystem Research Institute, Evanston, Illinois 60201, USA. [3]Department of Computer Science, Columbia University, New York, New York 10027, USA. [4]Medical Research Council-Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, UK. [5]MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, School of Medicine, Heath Park, Cardiff CF23 6BQ, UK. [6]Department of Health Research and Policy, Stanford University, Stanford, California 94304, USA. [7]Queensland Centre for Mental Health Research, and Queensland Institute for Medical Research, Brisbane, Queensland 4072, Australia. [8]Department of Psychiatry, University of Colorado Denver, Aurora, Colorado 80045, USA. [9]Department of Psychiatry and Behavioral Sciences, Atlanta Veterans Affairs Medical Center, and Emory University, Atlanta, Georgia 30322, USA. [10]Department of Psychiatry, Washington University, St Louis, Missouri 63110, USA. [11]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. [12]School of Nursing, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, USA. [13]Department of Psychiatry, University of California at San Francisco, San Francisco, California 94143, USA. [14]Mental Health Clinical Research Center, and Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, Iowa 52242, USA. [15]Department of Neurology, School of Medicine, University of California at San Francisco, San Francisco, California 94143, USA. [16]Center for Genotyping and Analysis, Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [17]Departments of Psychiatry, and Human Genetics, Virginia Commonwealth University, Richmond, Virginia 23298, USA.

©2009 Macmillan Publishers Limited. All rights reserved

Table 1 | MGS GWAS results

SNP	Chromosome/ band	Location (bp)	Odds ratio	*P* value	Gene(s)	Function/relevance
European-ancestry analysis						
rs13025591	2q37.2	236460082	1.225	4.59×10^{-7}	*CENTG2*	GTPase activator; deletions reported in autism cases[14]
rs16941261	15q25.3	86456524	1.255	8.10×10^{-7}	*NTRK3*	Tyrosine receptor kinase; MAPK signalling
rs10140896	14q31.3	88288291	1.216	9.49×10^{-7}	*EML5*	Microtubule assembly
rs17176973	5p15.2	10864474	1.679	2.16×10^{-6}		(50 kb upstream of *DAP*; mediates interferon-γ-induced cell death)
rs17833407	9p21.3	21738320	0.804	3.02×10^{-6}		(54 kb upstream of *MTAP*; enzyme involved in polyamine metabolism)
rs1635239	Xp22.33	3242699	0.790	3.04×10^{-6}	*MXRA5*	Cell adhesion protein
rs915071	14q12	31503609	0.834	3.94×10^{-6}		(102 kb downstream of *NUBPL*; nucleotide binding protein-like)
rs11061935	12p13.33	1684035	0.773	4.06×10^{-6}	*ADIPOR2*	Adiponectin (antidiabetic drug) receptor
rs6809315	3q13.11	107360155	0.828	7.58×10^{-6}		
rs1864744	14q31.3	88020759	0.828	7.59×10^{-6}	*PTPN21*	Regulation of cell growth and differentiation
rs1177749	10q23.33	97887981	0.835	1.29×10^{-5}	*ZNF518*	Regulation of transcription
rs17619975	6p22.3	15510731	0.611	1.49×10^{-5}	*JARID2*	Neural tube formation; histone demethylase; adjacent to *DTNBP1* (candidate gene)
African-American analysis						
rs1851196	2q34	212178865	0.733	2.14×10^{-6}	*ERBB4*	Neuregulin receptor
rs3751954	17q25.3	75368080	0.528	4.59×10^{-6}	*CBX2*	Polycomb protein; histone modifications, maintenance of transcriptional repression
rs10865802	3p24.2	25039902	1.330	6.73×10^{-6}		
rs17149424	9q34.13	134523705	1.680	8.00×10^{-6}	*DDX31*	RNA helicase family; embryogenesis
rs2162361	10q23.31	90037689	2.020	9.19×10^{-6}	*RNLS*	Degrades catecholamine (regulation of vascular tone)
rs17149524	9q34.13	134546628	1.642	9.59×10^{-6}	*GTF3C4*	Required for RNA polymerase III function
rs2587562	8q13.3	73153592	1.301	1.56×10^{-5}	*TRPA1*	Cannabinoid receptor (cannabis may ↑ schizophrenia risk[1]); pain, sound perception
rs4316112	8p12	32539889	0.564	1.59×10^{-5}	*NRG1*	Neuregulin 1; schizophrenia candidate gene; neuronal development
rs4732838	8p21.1	28098106	0.768	1.68×10^{-5}	*ELP3*	Histone acetyltransferase, RNA polymerase III elongator
rs9927946	16p13.2	8990868	0.718	1.70×10^{-5}		(26 kb upstream of *UPS7*; ubiquitin fusion protein cleavage; induction of apoptosis)
rs13065441	3q26.2	172478029	0.626	1.94×10^{-5}	*TNIK*	Stress-activated serine/threonine kinase
rs2729993	8p12	34116593	0.687	2.14×10^{-5}		

Shown are the top 12 results (excluding duplicates—SNPs in the same genes or regions with less significant results) of the MGS European-ancestry (2,681 cases, 2,653 controls) and African-American (1,286 cases, 973 controls) MGS GWAS analyses. Listed are genes within 10 kb of the SNP and annotated information on possible functional relevance; or (in parentheses), information on genes within 150 kb. The odds ratio is for the tested allele (see Supplementary Data 2). Supplementary Data 1 contains results for all SNPs with $P < 0.001$, and full gene names. Results of a further exploratory analysis that combined the two data sets are summarized in Supplementary Results 1, Supplementary Table 18 and Supplementary Data 1.

overlap is seen between European-ancestry and African-American samples. This could be because there are actually major differences between the sets of segregating common disease variants in these two populations, and/or because many risk variants are tagged by different GWAS markers or not adequately tagged by the GWAS array, which has poorer coverage of alleles that are more frequent in African populations. The hypothesis underlying our combined analysis, on the other hand, was that there could also be allelic effects common to these populations.

For many common diseases, common risk alleles with genotypic relative risks in the range of 1.1 to 1.2 have been detected when samples were combined to create much larger data sets[5]. Therefore, we carried out a meta-analysis of European-ancestry data with two other large studies: the ISC (3,322 cases, 3,587 controls) and the SGENE consortium (2,005 cases, 12,837 controls). Note that because the Aberdeen sample was part of both the ISC and SGENE consortia, Aberdeen data were excluded from SGENE association tests for the meta-analysis. To identify the regions containing the strongest findings across the three studies (which used several Affymetrix and Illumina genotyping platforms), each group created a list of the SNPs with the best *P* values in its final analysis (for example, those with $P < 0.001$ in MGS), and provided the other groups with its *P* values for the SNPs on their lists, on the basis of the genotyped or imputed data or data for the best proxy based on linkage disequilibrium. On the basis of these initial results, all available data for genotyped SNPs and imputed HapMap II SNPs were then shared for regions of interest, of which four emerged from the European-ancestry data: 1p21.3 (*PTBP2*), 4q33 (*NEK1*), 6p22.1–6p21.31 (extended major histocompatibility complex (MHC) region) and 18q21.2 (*TCF4*). We then combined *P* values for all SNPs in each region by appropriately weighting *Z* scores for sample size, accounting for the direction of association in each sample.

In the meta-analysis of European-ancestry MGS, ISC and SGENE data sets, seven SNPs on chromosome 6p22.1 yielded genome-wide significant evidence for association. These SNPs span 209 kilobases (kb) and are in strong linkage disequilibrium ($r^2 > 0.9$), with substantial linkage disequilibrium across 1.5 megabases (Mb) (Table 2 and Fig. 1). Because of the strong linkage disequilibrium among these SNPs, it is unclear whether the signal is driven by one or several genes, by intergenic elements, or by longer haplotypes that include susceptibility alleles in many genes. The region includes several types of genes of potential interest. The strongest evidence for association was observed in and near a cluster of histone protein genes, which could be relevant to schizophrenia through their roles in regulation of DNA transcription and repair[6,7] or their direct role in antimicrobial defence[8]. Other genes in the broad region are involved in chromatin structure (*HMGN4*), transcriptional regulation (*ABT1*, *ZNF322A* and *ZNF184*), immunity (*PRSS16*; the butyrophilins[9]), G-protein-coupled-receptor signalling (*FKSG83*) and in the nuclear pore complex (*POM121L2*), although the functions of many genes in the region (and of intergenic sequence variants) are not well understood.

P values of less than 10^{-7} were also observed in the meta-analysis in *HLA-DQA1* ($P = 6.88 \times 10^{-8}$, Table 2), suggesting autoimmune mechanisms. This gene is in the class II HLA region, which is not in linkage disequilibrium with 6p22.1 in the MGS sample. We note also that the MGS GWAS (see Supplementary Data 1, European-ancestry results) produced some evidence for association in the *FAM69A-EVI-RPL5* gene cluster, which has been implicated in multiple sclerosis, a DQA-associated autoimmune disorder[10].

Furthermore, in an analysis reported in the companion paper by the ISC[4], case-control status in the MGS sample could be predicted with very strong statistical significance on the basis of an aggregate test of large numbers of common alleles, weighted by their odds ratios in the single-SNP association analysis of the ISC sample (see ref. 4 for

©2009 Macmillan Publishers Limited. All rights reserved

Table 2 | Meta-analysis results in the extended MHC class I and II regions

				MGS freqs		*P* values				Odds ratios			Information				
rs ID	No.	bp	A1/2	Ctrl	Cases	Comb	MGS	ISC	SGENE	MGS	ISC	SGENE	MGS	ISC	SGENE	Band	Gene(s) (location)
rs6939997	1	25929203	T/C	0.089	0.080	4.90×10^{-7}	1.40×10^{-1}	5.66×10^{-4}	2.85×10^{-4}	0.900	0.785	0.712	0.978	0.851	0.955	6p22.2	*SLC17A1* (intr)
rs13199775	2	25936761	T/A	0.087	0.075	1.19×10^{-7}	5.12×10^{-2}	5.66×10^{-4}	2.57×10^{-4}	0.869	0.785	0.710	0.999	0.851	0.956	6p22.2	*SLC17A1* (intr)
rs9461219	3	25944906	G/C	0.087	0.075	4.72×10^{-7}	4.99×10^{-2}	2.68×10^{-3}	2.52×10^{-4}	0.868	0.826	0.710	1.000	1.000	0.956	6p22.2	*SLC17A1* (up), *SLC17A3* (dwn)
rs9467626	4	25981725	A/C	0.091	0.080	1.65×10^{-7}	5.32×10^{-2}	7.47×10^{-4}	2.68×10^{-4}	0.872	0.784	0.710	0.988	0.806	0.959	6p22.2	*SLC17A3* (intr)
rs2072806	5	26493072	G/C	0.115	0.097	9.27×10^{-7}	5.91×10^{-3}	2.80×10^{-4}	3.40×10^{-2}	0.836	0.814	0.857	0.959	1.000	0.971	6p22.1	*BTN3A2* (dwn), *BTN2A2* (intr)
rs2072803	6	26500494	C/G	0.115	0.097	8.19×10^{-7}	5.53×10^{-3}	2.80×10^{-4}	3.23×10^{-2}	0.835	0.814	0.856	0.960	1.000	0.974	6p22.1	*BTN2A2* (intr bound), *BTN3A1* (up)
rs6904071	7	27155235	A/G	0.186	0.166	$\mathbf{1.78\times10^{-8}}$	1.21×10^{-2}	3.03×10^{-4}	3.72×10^{-4}	0.879	0.819	0.799	0.987	0.851	0.988	6p22.1	
rs926300	8	27167422	T/A	0.186	0.166	$\mathbf{1.06\times10^{-8}}$	1.22×10^{-2}	3.03×10^{-4}	2.08×10^{-4}	0.879	0.819	0.791	0.989	0.851	0.988	6p22.1	
rs6913660	9	27199404	A/C	0.184	0.164	$\mathbf{2.36\times10^{-8}}$	1.71×10^{-2}	3.03×10^{-4}	3.35×10^{-4}	0.884	0.819	0.798	0.998	0.851	1.000	6p22.1	*HIST1H2AG* (up), *HIST1H2BJ* (dwn)
rs13219181	10	27244204	G/A	0.183	0.163	$\mathbf{1.29\times10^{-8}}$	1.47×10^{-2}	3.03×10^{-4}	2.05×10^{-4}	0.881	0.819	0.791	0.983	0.851	0.989	6p22.1	
rs13194053	11	27251862	C/T	0.182	0.162	$\mathbf{9.54\times10^{-9}}$	1.45×10^{-2}	3.03×10^{-4}	1.48×10^{-4}	0.880	0.819	0.783	0.977	0.851	0.978	6p22.1	
rs13219354	12	27293643	C/T	0.118	0.102	1.12×10^{-7}	3.59×10^{-2}	5.11×10^{-4}	4.39×10^{-4}	0.877	0.823	0.766	0.994	1.000	1.000	6p22.1	
rs3800307	13	27293771	A/T	0.205	0.183	$\mathbf{4.35\times10^{-8}}$	1.34×10^{-2}	3.35×10^{-3}	6.12×10^{-5}	0.886	0.880	0.787	1.000	1.000	0.988	6p22.1	
rs13212921	14	27313401	T/C	0.124	0.108	1.28×10^{-7}	3.09×10^{-2}	5.76×10^{-4}	5.53×10^{-4}	0.873	0.823	0.770	0.960	0.940	0.968	6p22.1	
rs4452638	15	27337244	A/G	0.117	0.102	2.68×10^{-7}	4.51×10^{-2}	3.96×10^{-4}	1.11×10^{-3}	0.882	0.821	0.780	1.000	1.000	0.987	6p22.1	*PRSS16* (dwn)
rs6938200	16	27339129	G/A	0.205	0.188	3.02×10^{-7}	5.28×10^{-2}	2.40×10^{-3}	1.51×10^{-4}	0.910	0.874	0.801	1.000	0.992	1.000	6p22.1	*PRSS16* (dwn)
rs6932590	17	27356910	C/T	0.240	0.215	7.13×10^{-8}	3.37×10^{-3}	2.23×10^{-3}	8.48×10^{-4}	0.872	0.877	0.834	0.973	0.933	1.000	6p22.1	
rs3800316	18	27364081	C/A	0.257	0.229	$\mathbf{3.81\times10^{-8}}$	7.19×10^{-4}	3.51×10^{-3}	1.09×10^{-3}	0.856	0.886	0.834	0.975	0.956	0.977	6p22.1	
rs7746199	19	27369303	T/C	0.185	0.160	5.03×10^{-8}	6.85×10^{-4}	8.77×10^{-4}	5.70×10^{-3}	0.837	0.859	0.839	1.000	1.000	1.000	6p22.1	
rs3800318	20	27371620	T/A	0.180	0.158	6.38×10^{-8}	2.80×10^{-3}	8.82×10^{-4}	2.27×10^{-3}	0.855	0.859	0.822	0.998	1.000	0.981	6p22.1	
rs16897515	21	27385999	A/C	0.173	0.154	1.83×10^{-7}	1.22×10^{-2}	6.40×10^{-4}	2.16×10^{-3}	0.876	0.853	0.816	1.000	1.000	0.972	6p22.1	*POM121L2* (missense)
rs13195040	22	27521903	G/A	0.087	0.076	2.50×10^{-7}	1.04×10^{-1}	3.00×10^{-5}	2.82×10^{-3}	0.892	0.714	0.768	1.000	1.000	0.983	6p22.1	*ZNF184* (dwn)
rs10484399	23	27642507	G/A	0.091	0.080	3.50×10^{-7}	1.09×10^{-1}	8.58×10^{-6}	8.69×10^{-3}	0.895	0.762	0.795	1.000	1.000	1.000	6p22.1	
rs17693963	24	27818144	C/A	0.101	0.085	2.81×10^{-7}	2.87×10^{-2}	6.00×10^{-5}	8.85×10^{-3}	0.864	0.781	0.795	1.000	0.993	0.986	6p22.1	
rs7776351	25	27834710	T/C	0.255	0.236	3.22×10^{-7}	2.83×10^{-2}	1.13×10^{-4}	6.51×10^{-3}	0.905	0.855	0.865	1.000	1.000	1.000	6p22.1	
rs12182446	26	27853717	A/G	0.247	0.227	4.77×10^{-7}	2.99×10^{-2}	7.17×10^{-5}	1.22×10^{-2}	0.905	0.850	0.874	0.999	1.000	0.990	6p22.1	
rs9272219		32710247	T/G	0.283	0.263	6.88×10^{-8}	1.32×10^{-2}	2.24×10^{-5}	1.03×10^{-2}	0.897	0.847	0.880	1.000	1.000	0.977	6p21.32	*HLA-DQA1* (up)
rs9272535		32714734	A/G	0.283	0.263	8.87×10^{-8}	1.63×10^{-2}	2.47×10^{-5}	9.92×10^{-3}	0.900	0.848	0.879	1.000	1.000	0.977	6p21.32	*HLA-DQA1* (intr)

No. refers to the numbering of SNPs in the linkage disequilibrium diagram in Fig. 1. A1/2, alleles 1 and 2 (minor/major, forward strand); comb, combined analysis (meta-analysis) of the three data sets; ctrl, controls; freqs, minor allele frequencies. For SNP locations (as reported by http://snpper.chip.org): dwn, downstream (within 10 kb of transcribed sequence); intr, intron; intr bound, intron–exon boundary; up, upstream (within 10 kb). Shown are data for all SNPs with $P<10^{-6}$ in the extended MHC class I and II regions of chromosome 6p, in the meta-analysis of European-ancestry MGS, ISC and SGENE data (see Table 1 and Supplementary Methods). Odds ratios are reported for the minor allele. Genome-wide significant meta-analysis *P* values ($P<5\times10^{-8}$) are shown in bold (see Supplementary Methods as regards this threshold). The information content for the SNP for each study is the imputation r^2 reported by MACH 1.0 for MGS, and the PLINK and IMPUTE information measure for ISC and SGENE, respectively. '1.000' indicates an SNP that was genotyped or had a perfect proxy. Genes within 10 kb of a SNP are listed; see Fig. 1 for more details for 6p22.1. See Supplementary Data 1 for further meta-analysis results.

©2009 Macmillan Publishers Limited. All rights reserved

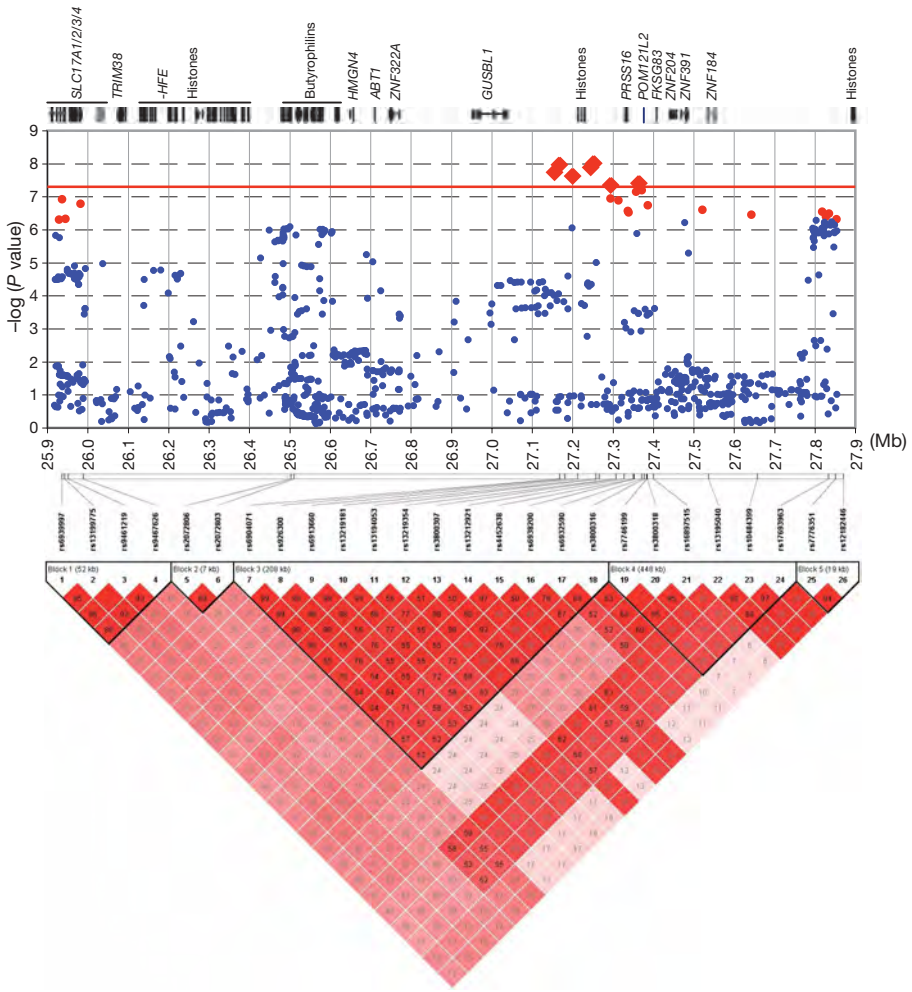


Figure 1 | Chromosome 6p22.1: genetic association and linkage disequilibrium results in European-ancestry samples. Genome-wide significant evidence for association ($P < 5 \times 10^{-8}$, threshold shown by solid red line, SNPs by large red diamonds) was observed at seven SNPs across 209 kb. *P* values are shown for all genotyped and imputed SNPs (25,900,000–27,875,000 bp) for the meta-analysis of European-ancestry MGS, ISC and SGENE samples (8,008 cases, 19,077 controls). Red circles indicate other SNPs with $P < 5 \times 10^{-7}$. Not shown are two SNPs in *HLA-DQA1* (6p21.32; lowest $P = 6.88 \times 10^{-8}$, 32,710,247 bp; see Supplementary Data 1). Locations are shown for RefSeq genes and *POM121L2*. Pairwise linkage disequilibrium relationships are shown for 26 SNPs with $P < 10^{-7}$ (except that SNPs 5 and 6 are shown, despite slightly larger *P* values, to illustrate linkage disequilibrium for that segment; and a SNP in strong linkage disequilibrium with SNPs 25 and 26 is omitted). Linkage disequilibrium was computed from MGS European-ancestry genotyped and imputed SNP data. The signal is poorly localized because of strong linkage disequilibrium: of the seven significant SNPs, 7–8 and 9–11 are in nearly perfect linkage disequilibrium; they are in or within ~30–50 kb of a cluster of five histone genes (*HIST1H2BJ*, *HIST1H2AG*, *HIST1H2BK*, *HIST1H4I* and *HIST1H2AH*; 27,208,073–27,223,325 bp). These SNPs are in moderately strong linkage disequilibrium ($r^2 = 0.52$–0.77) with two other significant SNPs 70–140 kb away, upstream of *PRSS16* (SNP 13) or between *PRSS16* and *POM121L2* (SNP 18). See Table 2 and Supplementary Figs 10 and 11 for further details.

details). As expected, results were similar for an analysis with MGS as the discovery sample and ISC as the target (see Supplementary Results 3). As discussed in the ISC paper, the results demonstrate that a substantial proportion of variance may be explained by many common variants, most of them with small effects that cannot be detected one at a time.

We have identified a region of association of common SNPs with schizophrenia on chromosome 6p22.1. Further research will be required to identify the sequence variation in this region that alters susceptibility, and the mechanisms by which this occurs. The results of this meta-analysis and of the aggregate analysis of multiple alleles reported in the ISC paper strongly suggest that individual common variants have small effects on schizophrenia risk, and that still larger samples may be valuable. The larger goal of research in the field will be to detect and understand the full range of rare and common sequence and structural schizophrenia susceptibility variants. Association findings will advance knowledge of pathophysiological mechanisms, even if they initially explain small proportions of genetic variance. Future advances in the knowledge of gene and protein functions and interactions should make it possible to dissect the functional sets of pathogenic variants on the basis of previous hypotheses.

METHODS SUMMARY

Details of MGS subject recruitment and sample characteristics are provided in the Supplementary Methods (section A1). DNA samples were genotyped using the Affymetrix 6.0 array at the Broad Institute. Samples (5.3%) were excluded for high missing data rates, outlier proportions of heterozygous genotypes, incorrect sex or genotypic relatedness to other subjects. SNPs (7% for African-American, 25% for European-ancestry and 27% for combined analyses) were excluded for minor allele frequencies less than 1%, high missing data rates, Hardy–Weinberg deviation (controls), or excessive Mendelian errors (trios), discordant genotypes (duplicate samples) or large allele frequency differences among DNA plates. Principal component scores reflecting continental and within-Europe ancestries of each subject were computed and outliers were excluded. Genomic control λ values for autosomes after quality control procedures were 1.042 for African-American and 1.087 for the larger European-ancestry and combined analyses.

For MGS, association of single SNPs to schizophrenia was tested by logistic regression (trend test) using PLINK[11], separately for European-ancestry, African-American and combined data sets, correcting for principal component scores that reflected

©2009 Macmillan Publishers Limited. All rights reserved

geographical gradients or that differed between cases and controls, and for sex for chromosome X and pseudoautosomal SNPs. Genotypic data were imputed for 192 regions surrounding the best findings, and for further regions selected for meta-analysis[12]. Detailed results are available in Supplementary Data 1 and 2, and complete results are available from dbGAP (www.ncbi.nlm.nih.gov/sites/entrez?db=gap).

Meta-analysis of the MGS, ISC and SGENE data sets was carried out by combining *P* values for all SNPs (in the selected regions) for which genotyped or imputed data were available for all data sets, with weights computed from case-control sample sizes. See the companion papers for details of the ISC and SGENE analyses[4,13].

Received 29 May; accepted 10 June 2009.
Published online 1 July; corrected 6 August 2009 (see full-text HTML version for details).

1. Tandon, R., Keshavan, M. S. & Nasrallah, H. A. Schizophrenia, ''just the facts'' what we know in 2008. 2. Epidemiology and etiology. *Schizophr. Res.* **102,** 1–18 (2008).
2. Cook, E. H. Jr & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455,** 919–923 (2008).
3. Arnold, S. E., Talbot, K. & Hahn, C. G. Neurodevelopment, neuroplasticity, and new genes for schizophrenia. *Prog. Brain Res.* **147,** 319–345 (2005).
4. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* doi:10.1038/nature08185 (this issue).
5. Manolio, T. A., Brooks, L. D. & Collins, F. S. A. HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
6. Adegbola, A., Gao, H., Sommer, S. & Browning, M. A novel mutation in *JARID1C/SMCX* in a patient with autism spectrum disorder (ASD). *Am. J. Med. Genet. A.* **146A,** 505–511 (2008).
7. Costa, E. *et al.* Reviewing the role of DNA (cytosine-5) methyltransferase overexpression in the cortical GABAergic dysfunction associated with psychosis vulnerability. *Epigenetics* **2,** 29–36 (2007).
8. Kawasaki, H. & Iwamuro, S. Potential roles of histones in host defense as antimicrobial agents. *Infect. Disord. Drug Targets* **8,** 195–205 (2008).
9. Malcherek, G. *et al.* The B7 homolog butyrophilin BTN2A1 is a novel ligand for DC-SIGN. *J. Immunol.* **179,** 3804–3811 (2007).
10. Oksenberg, J. R., Baranzini, S. E., Sawcer, S. & Hauser, S. L. The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nature Rev. Genet.* **9,** 516–526 (2008).
11. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
12. Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* **84,** 235–250 (2009).
13. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* doi:10.1038/nature08186 (this issue).
14. Wassink, T. H. *et al.* Evaluation of the chromosome 2q37.3 gene *CENTG2* as an autism susceptibility gene. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **136B,** 36–44 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the study participants, and the research staff at the study sites. This study was supported by funding from the National Institute of Mental Health (USA) and the National Alliance for Research on Schizophrenia and Depression. Genotyping of part of the sample was supported by the Genetic Association Information Network (GAIN), and by The Paul Michael Donovan Charitable Foundation. Genotyping was carried out by the Center for Genotyping and Analysis at the Broad Institute of Harvard and MIT with support from the National Center for Research Resources (USA). The GAIN quality control team (G. R. Abecasis and J. Paschall) made important contributions to the project. We thank S. Purcell for assistance with PLINK.

Author Contributions J.S., D.F.L. and P.V.G. wrote the first draft of the paper. P.V.G., D.F.L., A.R.S., B.J.M., A.O., F.A., C.R.C., J.M.S., N.G.B., W.F.B., D.W.B., R.R.C. and R.F. oversaw the recruitment and clinical assessment of MGS participants, and the clinical aspects of the project and analysis. A.R.S., D.F.L. and P.V.G. performed database curation. D.F.L., J.S., I.P., F.D., P.A.H., A.S.W. and P.V.G. designed the analytical strategy and analysed the data. D.B.M. oversaw the Affymetrix 6.0 genotyping, and J.D., Y.Z., A.R.S. and P.V.G. performed the preparative genotyping and experimental work. J.R.O. contributed to the interpretation of data in the MHC/HLA region, and K.S.K. contributed to the approach to clinical data. P.V.G. coordinated the overall study. All authors contributed to the current version of the paper.

Author Information Data have been deposited at dbGaP (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) under accessions phs000021.v2.p1 and phs000167.v1.p1, and the NIMH Center for Collaborative Genetic Studies on Mental Disorders (http://www.nimhgenetics.org) under studies 6, 29 and 29C. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.V.G. (pgejman@mac.com).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 460|6 August 2009|doi:10.1038/nature08167

LETTERS

Switch in FGF signalling initiates glial differentiation in the *Drosophila* eye

Sigrídur Rut Franzdóttir[1]*†, Daniel Engelen[1]*, Yeliz Yuva-Aydemir[1], Imke Schmidt[1], Annukka Aho[1] & Christian Klämbt[1]

The formation of a complex nervous system requires the intricate interaction of neurons and glial cells. Glial cells generally migrate over long distances before they initiate their differentiation, which leads to wrapping and insulation of axonal processes[1,2]. The molecular pathways coordinating the switch from glial migration to glial differentiation are largely unknown[3]. Here we demonstrate that, within the *Drosophila* eye imaginal disc, fibroblast growth factor (FGF) signalling coordinates glial proliferation, migration and subsequent axonal wrapping. Glial differentiation in the *Drosophila* eye disc requires a succession from glia–glia interaction to glia–neuron interaction[4]. The neuronal component of the fly eye develops in the peripheral nervous system within the eye–antennal imaginal disc, whereas glial cells originate from a pool of central-nervous-system-derived progenitors and migrate onto the eye imaginal disc[5–8]. Initially, glial-derived Pyramus, an FGF8-like ligand, modulates glial cell number and motility. A switch to neuronally expressed Thisbe, a second FGF8-like ligand, then induces glial differentiation. This switch is accompanied by an alteration in the intracellular signalling pathway through which the FGF receptor channels information into the cell. Our findings reveal how a switch from glia–glia interactions to glia–neuron interactions can trigger formation of glial membrane around axonal trajectories. These results disclose an evolutionarily conserved control mechanism of axonal wrapping[2], indicating that *Drosophila* might serve as a model to understand glial disorders in humans.

Migration of glial cells onto the eye disc relies on unique homotypic cell interactions[4]. Glial cells are generated in the perineurium and migrate along specialized glial cells, the subperineurium, onto the eye disc (Fig. 1a–c). Perineurial glial cells contact nascent photoreceptor axons only when they reach the edge of the growing eye field[4]. On contact, glial cells switch behaviour and begin to differentiate, leading to a complete wrapping of ommatidial axon fascicles (Fig. 1b, c).

How are these different aspects of gliogenesis controlled during development? The *Drosophila* genome harbours two FGF-receptor genes[9,10]. The messenger RNA of the FGF receptor *heartless* (*htl*) is expressed in the eye disc glia with most prominent expression at the front of the migratory glial cell population (Fig. 1d). Htl protein is expressed broadly in the glia and decorates glial projections following the photoreceptor axons (Fig. 1h–k and Supplementary Fig. 1). When we reduced specifically *htl* function in glial cells by expressing a dominant-negative form of the FGF receptor or by RNA-mediated interference (RNAi), we noted a 40% reduction in glial cell number, impaired migration and a lack of differentiation (Figs 1m, n, 3 and Supplementary Figs 2 and 3). Moreover, overexpression of a constitutively active Htl (λ-Htl)(ref. 11) resulted in an eightfold

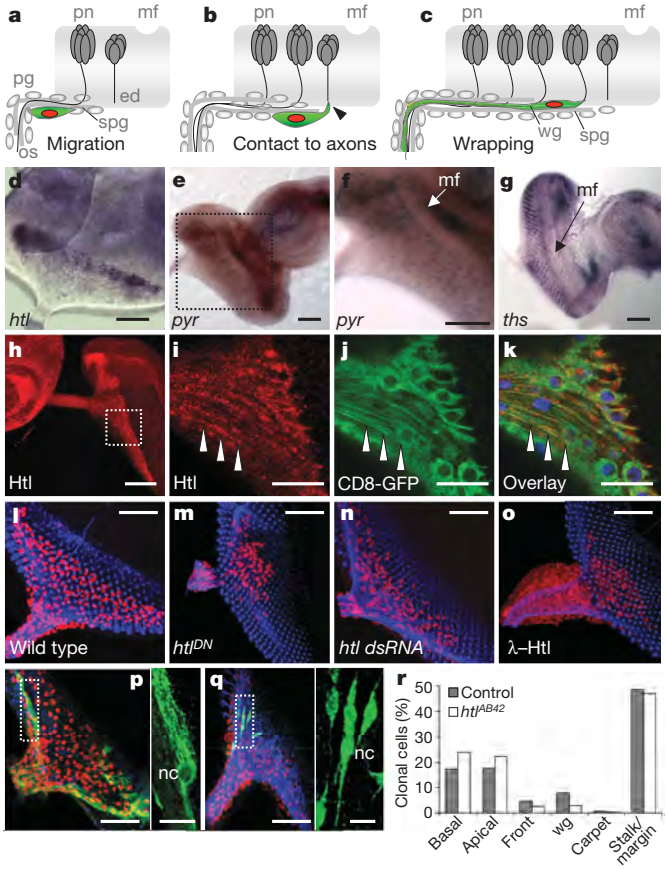


Figure 1 | The role of Htl during glial development in the eye disc. **a–c**, *Drosophila* eye disc glia development. ed, eye disc; mf, morphogenetic furrow; os, optic stalk; pg, perineurial glia; pn, photoreceptor neuron; spg, subperineurial glia; wg, wrapping glia. Glia–glia interactions accompany proliferation and migration (**a**); neuron–glia interactions (**b**, arrowhead) initiate glial wrapping (**c**). **d–g**, RNA *in situ* hybridization. **d**, *htl* mRNA. **e**, **f**, *pyr* mRNA is expressed anterior to the morphogenetic furrow (mf) and in glial cells posterior to the morphogenetic furrow (detail in **f**). **g**, *ths* mRNA. **h–k**, Confocal images of eye discs stained for Htl (red) and glial morphology (UAS-CD8-GFP expression driven by *repo-GAL4* (green) and Repo (blue)). **i–k**, Boxed area from **h**. Arrowheads indicate Htl expression in glial processes. **l**, Wild type, glial nuclei (red), differentiating photoreceptors (horseradish peroxidase, blue). Glial expression of a dominant-negative form of *htl* (htl^{DN}) (**m**) or *htl* dsRNA (**n**) is shown. **o**, Expression of activated Htl (λ-Htl). **p–r**, MARCM analysis. Homozygous control (**p**) or mutant htl^{AB42} clone (**q**). nc, nucleus. **r**, Quantification ($n = 536$ cells). wg, wrapping glia. Scale bars: **d–q**, 50 μm; amplified sections of **p** and **q**, 10 μm.

[1]Institut für Neurobiologie, Universität Münster, Badestr. 9, D-48149 Münster, Germany. †Present address: Biomedical Center, University of Iceland, 101 Reykjavik, Iceland.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

increase in glial cell number and also impaired glial migration (Fig. 1o and Supplementary Fig. 2). This demonstrates a central role of FGF signalling during gliogenesis in the *Drosophila* eye.

To dissect the requirement of *htl*, we generated small labelled glial cell clones homozygous for the htl^{AB42} null allele[12,13]. *htl*-deficient cell clones were only slightly smaller than wild-type clones and mutant cells migrated normally, enabling us to decipher the role of Htl during glial differentiation. The mutant cell clones comprise all glial cell types. However, instead of cells adopting the typical wrapping phenotype, thin, spindle-shaped cells are found (Fig. 1p–r). Thus, Htl seems to cell-autonomously control differentiation of the terminal glial cell type, the wrapping glia.

We addressed how FGF signalling can evoke seemingly different aspects of gliogenesis such as division, migration and axonal wrapping by analysing the two FGF8-like ligands, Pyramus (Pyr) and Thisbe (Ths), both of which activate Htl (refs 14–16). *pyr* is expressed in the eye disc anterior to the morphogenetic furrow and in the glia (Fig. 1e, f). *ths* is expressed only in photoreceptor neurons (Fig. 1g). This differential expression was verified by polymerase chain reaction with reverse transcription (RT–PCR) on cell-type-specific mRNA (Supplementary Fig. 4b). To determine the function of the two FGF8-like ligands, we used P-element insertions (Supplementary Fig. 4a). The hypomorphic allele $pyr^{c02915a}$ leads to a 50% reduction in glial cell number (Fig. 2d–f). An even stronger reduction in glial cell number was observed after pan-glial RNAi of *pyr* (Supplementary Fig. 5j–l). In contrast, reduced *pyr* function in the entire eye disc did not affect cell number, but caused irregular glial migration (Supplementary Fig. 5e–i). Overexpression of Pyr in glia or in the eye disc efficiently induced glial cell division and stimulated glial motility (Fig. 2g–k and Supplementary Movie 1). To test further the role of Pyr in glial migration, we generated small patches of cells expressing Pyr ahead of the morphogenetic furrow (Fig. 2l, m and Supplementary Movie 2). Glial cells can migrate along such ectopic Pyr-expressing cells and are able to cross the morphogenetic furrow, which they never do in wild type. Similarly, Pyr-expressing clones in the apical peripodial membrane ectopically attract glial cells to the apical side of the eye disc (Fig. 2 and Supplementary Movie 2). When Pyr-expressing clones were well anterior to the morphogenetic furrow, glial cells did not move towards this source of FGF8. In conclusion, glial cell number is mainly regulated by glia–glia interactions. Pyr initially acts as an auto- or paracrine signal regulating glial cell number, and subsequently facilitates glial migration.

One hypothesis on what stops glial migration and initiates the differentiation into wrapping glia is that neuronal Ths induces a switch in the developmental program of the eye glia. Indeed, loss of ths caused a glial overmigration phenotype, suggesting that ths participates in stopping glial migration (Supplementary Fig. 5m–o). Moreover, reduction of neuronal Ths expression resulted in severe differentiation phenotypes. Photoreceptor axons form abnormal fascicles and are associated with less wrapping glia processes (Fig. 3d, i). Comparable phenotypes were found in *ths* mutant larvae (Fig. 3m, n) or after suppression of *htl* function in the glia (Fig. 3c, f). Given that loss of neuronal Ths decreased axonal wrapping, increased neuronal expression of Ths is expected to promote axonal wrapping, similar to what is described for the neuregulin/Erb-B system controlling myelination in the mammalian peripheral nervous system[2,17]. Indeed, neuronal overexpression of Ths caused increased glial cell membrane formation (Fig. 3e, g). In wild-type larvae, the eight ommatidial axons are wrapped as one fascicle with a central axon that is rarely reached by glial processes (1 out of 173 ommatidia). In the presence of additional axonal Ths, more glial processes ensheath the fascicle and invade the ommatidial fascicle to contact the central axon (Fig. 3e, g; 20 out of 72 ommatidia). Likewise, expression of activated λ-Htl in the wrapping glia caused extensive hyperwrapping (Fig. 3h; 47 out of 129 ommatidia). This is most clearly seen at the example of the 12 axons of Bolwig's nerve, which normally are ensheathed as one single unit[18]. Neuronal expression of Ths leads to a growth of glial membranes around individual Bolwig axons. Further increase of FGF-receptor activity by expressing activated λ-Htl in the wrapping glia results in the isolation of every single axon by a glial sheath (Fig. 3j–l). Thus, the level of FGFR activity determines the extent of glial wrapping. Both FGF8 molecules can cause axonal hyperwrapping (Fig. 3g, p and Supplementary Fig. 7a, b) and can in principle trigger the same biological response, because neuronal expression of Pyr in a *ths* mutant eye disc rescues the glial wrapping phenotype (Fig. 3m–p and Supplementary Fig. 4a).

FGF-receptor activity controls both glia–glia as well as glia–neuron interactions. The switch between the two modalities is reflected by a change in the activating ligand. Initially, Pyr controls glial division and migration. Subsequently, Ths induces cell differentiation by controlling the degree of axonal wrapping. The switch in cellular response to FGF-receptor activity may be induced by different downstream signalling components and different signalling strength or duration. The canonical

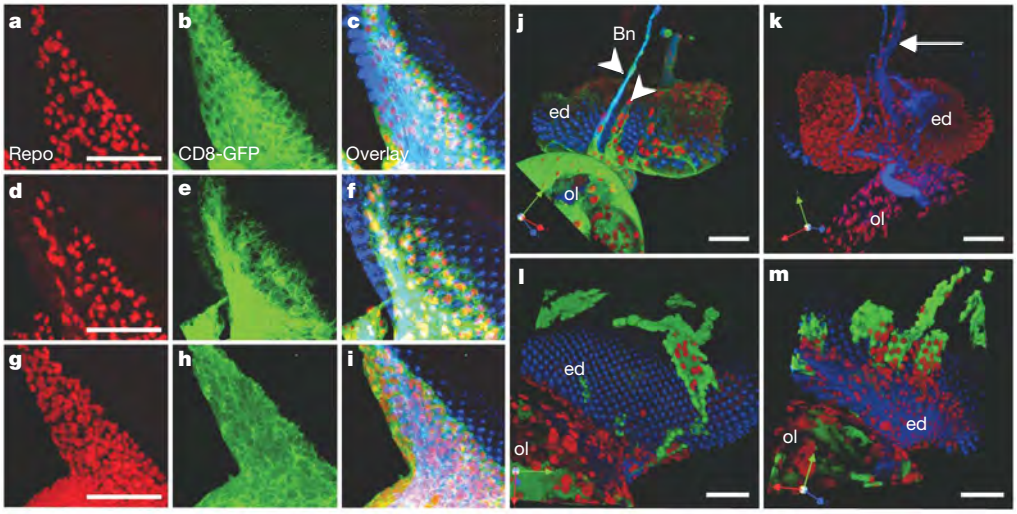


Figure 2 | Pyr directs glial proliferation and migration. Eye imaginal discs. Glial nuclei (red), neurons (blue) and glial outlines (green, UAS-CD8-GFP expression driven by *repo-GAL4* (**b**, **c**, **e**, **f**, **h**, **i**) or *repo-lexA lexAop-CD2-GFP* (**j**, **l**, **m**)) are shown. **a**–**c**, Wild type. **d**–**f**, $pyr^{c02915a}$ with 103 glial cells versus 210 in the wild type, $n = 6$. **g**–**i**, Glial overexpression of Pyr. **j**, **k**, Three-dimensional reconstruction of an *ey-GAL4* driven *P(XP)d06722* eye disc, apical (**j**) and basal (**k**). Glial cells accumulate along the peripodial membrane or Bolwig's nerve (Bn; arrowheads), or along ectopic axon projections (arrow). **l**, **m**, Glial cells migrate on Pyr-expressing cells in the apical (**l**) as well as in the basal part (**m**) of the eye disc. ed, eye disc; ol, optic lobe. See also Supplementary Movies 1 and 2. Scale bars, 50 μm.

©2009 Macmillan Publishers Limited. All rights reserved

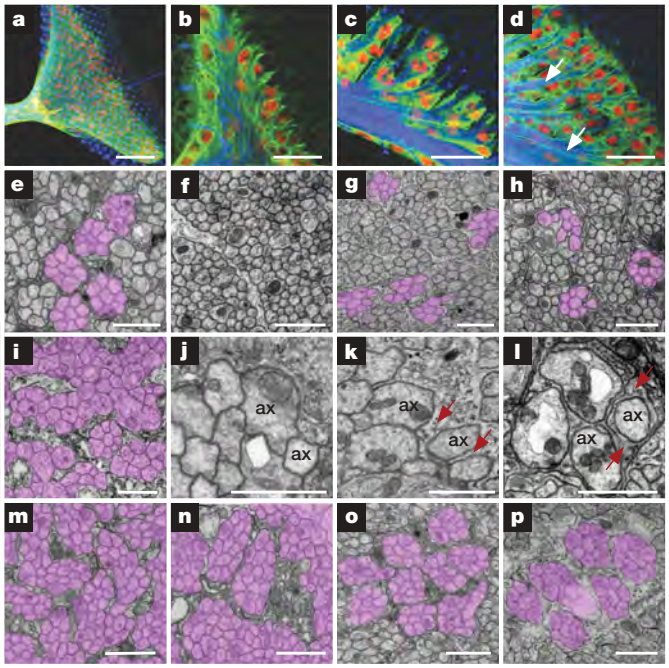


Figure 3 | Ths controls glial differentiation. **a**, **b**, **e**, Wild-type glia separate and ensheath eight axons from individual ommatidia. **c**, **d**, **f**, **i**, Glial *htl* dsRNA expression (**c**, **f**) and neuronal expression of *ths* dsRNA (**d**, **i**) suppress wrapping (arrows in **d**, purple in **i**). **g**, Expression of Ths in photoreceptor neurons or the expression of λ-Htl in wrapping glia (**h**) increase axonal wrapping. **j**, Axons from Bolwig's organ are not wrapped in the wild type. ax, axonal profile. **k**, **l**, Enhanced neuronal Ths expression (**k**) or expression of λ-Htl in wrapping glia (**l**) increases wrapping. **m**, Homozygous ths^{e02026} glia fails to wrap individual axon fascicles. **n**, ths^{e02026} /*D(2R)BSC25* glia shows stronger wrapping defects. **o**, In $pyr^{XP06722}$ ths^{e02026} /ths^{e02026}; *LGMR-GAL4* animals the wrapping defect is rescued. **p**, In $pyr^{XP06722}$ *LGMR-GAL4* animals ectopic expression of Pyr causes hyperwrapping. Scale bars: **a**, 50 μm; **b**–**d**, 20 μm; **e**–**p**, 1 μm.

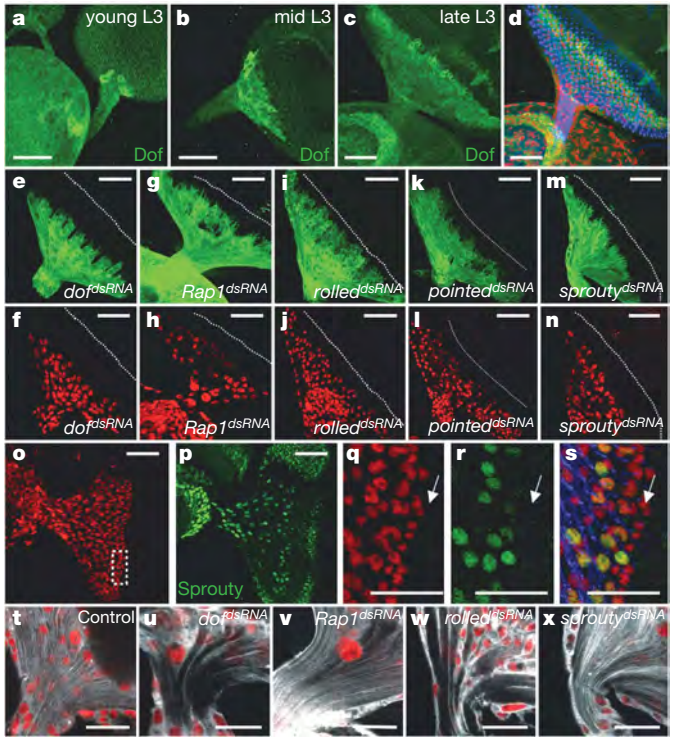


Figure 4 | FGF signalling in migrating and differentiating glial cells. **a**–**d**, Dof expression in glia (green). **e**–**n**, Glia-specific RNAi against the genotype indicated. Dotted lines indicate morphogenetic furrow. Glial cell bodies are green (UAS-CD8-GFP expression driven by *repo-GAL4*) and glial nuclei are red. Average cell numbers in comparison with wild type ($n = 6$): *dof*, 130 of 270; *Rap1*, 125 of 270; *rolled*, 221 of 240; *pointed*, 265 of 270; *sprouty*, 349 of 330. **o**–**s**, Expression of *sprouty* (*lacZ* insertion $P(PZ)eIF5B^{09143}$, green). **q**–**s**, Magnification of the region marked in **o** (Repo, red); **s**, overlay. The *sprouty* reporter is expressed in differentiating glia one row behind the front (arrow). **t**–**x**, Single confocal sections of the posterior eye disc and optic stalk. Glial cytoplasm is marked with UAS-CD8-GFP expression driven by *repo-GAL4* (grey), the gene affected is indicated. Scale bars: **a**–**p**, 50 μm; **q**–**x**, 25 μm.

FGF-receptor signalling cascade uses the adaptor protein Downstream of FGF receptor (Dof; also called Stumps) in *Drosophila* and feeds into the Ras mitogen-activated protein kinase (MAPK) cascade. Dof shows a pan-glial expression in the eye disc and is upregulated at the front of the migratory field (Fig. 4a–d). Block of *dof* function by glia-specific RNAi results in phenotypes similar to those observed after reduction of *htl* function (Fig. 4e, f, u; 52% reduction in glial cell number). Likewise, wrapping glia is missing in *dof* MARCM (mosaic analysis with a repressible cell marker) clones (Supplementary Fig. 6). Dof signals through small G proteins of the Ras family into the MAPK pathway to modulate the activity of transcriptional regulators[19]. *Drosophila* harbours three *Ras* genes. The knockdown of *Ras85D* and *Ras64B* leads to only subtle phenotypes. The Ras family member Rap1 mediates sustained MAPK activation and can act independently of Ras to control cell division and migration[20–22]. Glial-specific *Rap1* RNAi interferes with glial cell division and migration, but glial differentiation still proceeds relatively normally (Fig. 4g, h, v; 54% reduction in glial cell number). In contrast, knockdown of the Rolled MAPK results in lack of glial differentiation, but only mildly affects glial division or migration (Fig. 4i, j, w). A similar independency of Rap1 and Rolled has been reported[23,24]. Finally, RNAi pan-glial reduction of *pointed*, which encodes an ETS domain transcription factor downstream of FGF-receptor signalling, impairs glial differentiation (Fig. 4k, l). Likewise, MARCM analysis demonstrated that *pointed* is required for glial cell differentiation (Supplementary Fig. 6), supporting the notion that different signalling components accompany the altered cellular responses to FGFR activity.

In addition, negative regulators such as Sprouty are known to modulate FGF signalling at multiple steps[19,25]. *sprouty* expression is induced in wrapping glial cells shortly after their first contact with axons (Fig. 4o–s and Supplementary Fig. 8a–c). This induction depends only in part on FGF-receptor signalling. Expression of activated λ-Htl can induce *sprouty* expression in wrapping glia, but not in perineurial glial cells (Supplementary Fig. 8d–i). Thus, additional neuronal signals seem to be required to induce *sprouty* expression in wrapping glia. RNAi-mediated reduction of *sprouty* in wrapping glia results in excessive axonal wrapping similar to that observed after FGF-receptor activation (Figs 3g, h, 4x and Supplementary Fig. 7c, c′). Enhanced *sprouty* expression in the wrapping glia inhibits differentiation (Supplementary Fig. 7d, d′), indicating that *sprouty* is required to titrate FGF-receptor activity level during glial wrapping.

FGF signalling initially modulates proliferation and glial migration onto the eye disc, and subsequently neuronally expressed ths regulates the wrapping of axons. Unlike other developmental models[26,27], glial development is not controlled by graded FGF activation, but rather through stepwise regulation of FGF signalling. This is accompanied by a change in the FGF-receptor signalling intensity and the downstream transduction cascades. Moreover, the negative regulator Sprouty is expressed specifically in wrapping glia and fine-tunes FGFR activity to ensure correct axonal wrapping. Because myelin formation in mammals is also controlled by fine-tuning of receptor tyrosine kinase activity, the regulatory mechanisms underlying glial cell differentiation may be conserved[17,28]. Interestingly, whereas epidermal growth factor receptor activity is required for myelination in the peripheral nervous system and can also evoke myelin formation

©2009 Macmillan Publishers Limited. All rights reserved

in the central nervous system (CNS)[1,2], loss of epidermal growth factor receptor activity does not affect myelin formation in the CNS, leaving open the identification of the relevant receptor tyrosine kinase[29].

METHODS SUMMARY

Histology. Eye imaginal discs (>20 discs per genotype) were stained according to standard protocols. All antibodies used are listed in Methods. Specimens were analysed on a Zeiss 510 LSM confocal microscope, three-dimensional reconstructions and cell counts were made with Volocity (Improvision). For cell counts, five to six discs were used for each genotype/age. *In situ* hybridization was performed according to standard protocols. Electron microscopic analysis was performed as previously described[4].

Molecular biology. Anti-FGFR antibodies were induced in guinea-pigs (Eurogentec). Cell-type-specific mRNA was isolated as described[30].

Fly genetics. All fly stocks used are listed in Methods. Glial MARCM clones were generated with a glial-expressed *Flp* source and clonal patches of GAL4 expression were induced in a flip-out approach (see Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 April; accepted 28 May 2009.
Published online 13 July 2009.

1. Sherman, D. L. & Brophy, P. J. Mechanisms of axon ensheathment and myelin growth. *Nature Rev. Neurosci.* **6,** 683–690 (2005).
2. Nave, K. A. & Salzer, J. L. Axonal regulation of myelination by neuregulin 1. *Curr. Opin. Neurobiol.* **16,** 492–500 (2006).
3. Cafferty, P. & Auld, V. J. No pun intended: future directions in invertebrate glial cell migration studies. *Neuron Glia Biol.* **3,** 45–54 (2007).
4. Silies, M. *et al.* Glial cell migration in the eye disc. *J. Neurosci.* **27,** 13130–13139 (2007).
5. Wolff, T. & Ready, D. F. in *The Development of* Drosophila melanogaster (eds Bate, M. & Martinez Arias, A.) Ch. 22, 1277–1325 (Cold Spring Harbor Press, 1993).
6. Choi, K. W. & Benzer, S. Migration of glia along photoreceptor axons in the developing *Drosophila* eye. *Neuron* **12,** 423–431 (1994).
7. Hummel, T., Attix, S., Gunning, D. & Zipursky, S. L. Temporal control of glial cell migration in the *Drosophila* eye requires *gilgamesh*, *hedgehog*, and eye specification genes. *Neuron* **33,** 193–203 (2002).
8. Rangarajan, R., Gong, Q. & Gaul, U. Migration and function of glia in the developing *Drosophila* eye. *Development* **126,** 3285–3292 (1999).
9. Shishido, E., Ono, N., Kojima, T. & Saigo, K. Requirements of DFR1/Heartless, a mesoderm-specific *Drosophila* FGF-receptor, for the formation of heart, visceral and somatic muscles, and ensheathing of longitudinal axon tracts in CNS. *Development* **124,** 2119–2128 (1997).
10. Klambt, C., Glazer, L. & Shilo, B. Z. *breathless*, a *Drosophila* FGF receptor homolog, is essential for migration of tracheal and specific midline glial cells. *Genes Dev.* **6,** 1668–1678 (1992).
11. Michelson, A. M., Gisselbrecht, S., Buff, E. & Skeath, J. B. Heartbroken is a specific downstream mediator of FGF receptor signalling in *Drosophila*. *Development* **125,** 4379–4389 (1998).
12. Gisselbrecht, S., Skeath, J. B., Doe, C. Q. & Michelson, A. M. *heartless* encodes a fibroblast growth factor receptor (DFR1/DFGF-R2) involved in the directional migration of early mesodermal cells in the *Drosophila* embryo. *Genes Dev.* **10,** 3003–3017 (1996).
13. Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **22,** 451–461 (1999).
14. Gryzik, T. & Muller, H. A. *FGF8-like1* and *FGF8-like2* encode putative ligands of the FGF receptor Htl and are required for mesoderm migration in the *Drosophila* gastrula. *Curr. Biol.* **14,** 659–667 (2004).
15. Stathopoulos, A., Tam, B., Ronshaugen, M., Frasch, M. & Levine, M. *pyramus* and *thisbe*: FGF genes that pattern the mesoderm of *Drosophila* embryos. *Genes Dev.* **18,** 687–699 (2004).
16. Kadam, S., McMahon, A., Tzou, P. & Stathopoulos, A. FGF ligands in *Drosophila* have distinct activities required to support cell migration and differentiation. *Development* **136,** 739–747 (2009).
17. Michailov, G. V. *et al.* Axonal neuregulin-1 regulates myelin sheath thickness. *Science* **304,** 700–703 (2004).
18. Meinertzhagen, I. A. & Hanson, T. E. in *The Development of* Drosophila melanogaster (eds Bate, M. & Martinez Arias, A.) Ch. 24, 1363–1491 (Cold Spring Harbor Laboratory Press, 1993).
19. Thisse, B. & Thisse, C. Functions and regulations of fibroblast growth factor signaling during embryonic development. *Dev. Biol.* **287,** 390–402 (2005).
20. Mishra, S., Smolik, S. M., Forte, M. A. & Stork, P. J. Ras-independent activation of ERK signaling via the torso receptor tyrosine kinase is mediated by Rap1. *Curr. Biol.* **15,** 366–370 (2005).
21. York, R. D. *et al.* Rap1 mediates sustained MAP kinase activation induced by nerve growth factor. *Nature* **392,** 622–626 (1998).
22. Kooistra, M. R., Dube, N. & Bos, J. L. Rap1: a key regulator in cell-cell junction formation. *J. Cell Sci.* **120,** 17–22 (2007).
23. Asha, H., de Ruiter, N. D., Wang, M. G. & Hariharan, I. K. The Rap1 GTPase functions as a regulator of morphogenesis *in vivo*. *EMBO J.* **18,** 605–615 (1999).
24. Ishimaru, S., Williams, R., Clark, E., Hanafusa, H. & Gaul, U. Activation of the *Drosophila* C3G leads to cell fate changes and overproliferation during development, mediated by the RAS-MAPK pathway and RAP1. *EMBO J.* **18,** 145–155 (1999).
25. Mason, J. M., Morrison, D. J., Basson, M. A. & Licht, J. D. Sprouty proteins: multifaceted negative-feedback regulators of receptor tyrosine kinase signaling. *Trends Cell Biol.* **16,** 45–54 (2006).
26. Delfini, M. C., Dubrulle, J., Malapert, P., Chal, J. & Pourquie, O. Control of the segmentation process by graded MAPK/ERK activation in the chick embryo. *Proc. Natl Acad. Sci. USA* **102,** 11343–11348 (2005).
27. Dubrulle, J. & Pourquie, O. *fgf8* mRNA decay establishes a gradient that couples axial elongation to patterning in the vertebrate embryo. *Nature* **427,** 419–422 (2004).
28. Edenfeld, G. *et al.* The splicing factor Crooked neck associates with the RNA-binding protein HOW to control glial cell maturation in *Drosophila*. *Neuron* **52,** 969–980 (2006).
29. Brinkmann, B. G. *et al.* Neuregulin-1/ErbB signaling serves distinct functions in myelination of the peripheral and central nervous system. *Neuron* **59,** 581–595 (2008).
30. Stork, T. *et al. Drosophila* neurexin IV stabilizes neuron-glia interactions at the CNS midline by binding to Wrapper. *Development* **136,** 1251–1261 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Müller for providing flies and advice, M. Leptin, M. Freeman, the Bloomington stock centre and the Vienna Drosophila RNAi Center (VDRC) for sending flies and antibodies, H. Aberle, S. Bogdan, T. Hummel, E. Raz, M. Silies and R. Stephan for discussions. The work was funded through grants of the Deutsche Forschungsgemeinschaft.

Author Contributions S.R.F. performed the analysis of FGFR function during eye disc glial development (MARCM analysis, RNA interference, gain of function analysis and confocal studies), generated the antibodies and devised the genetic crosses. D.E. conducted the electron microscopic analysis and determined the phenotype of *pyramus* and *thisbe* mutants. Y.Y.-A. performed the cell-type-specific mRNA isolation, I.S. conducted the *pnt* MARCM analysis and A.A. contributed to the statistical analysis. C.K. contributed ideas, coordinated the work and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.K. (klaembt@uni-muenster.de).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Histology, electron microscopic analysis, quantification. Eye imaginal discs were stained according to standard protocols. We used the following antibodies: mouse anti-Repo (Developmental Studies Hybridoma Bank), rabbit anti-Dof (M. Leptin), rabbit anti-β-galactosidase (Cappel), rabbit anti-GFP (Invitrogen) and anti-HRP Cy5 (Dianova). Secondary antibodies were from Dianova. Specimens were analysed on a Zeiss 510 LSM confocal microscope, orthogonal sections and overlays were made using Zeiss software. 3D reconstructions were made with Volocity 4.0 (Improvision). The classifier tool was set in Volocity Quantification with an intensity range from 60 to 255, 25 μm^3 were defined as minimal countable object. Cell numbers were obtained from at least five imaginal discs and were compared with wild type of the comparable age (number of neuronal cell rows). *pyramus* mutant larvae die during the early L3 stage, at which point the imaginal disc usually has 10–12 rows of photoreceptor cells. *In situ* hybridization was performed according to standard protocols. Probes were generated by DIG RNA labelling (Roche). Electron microscopic analysis was performed as previously described[4].

Molecular biology. Part of the *htl* coding sequence (802 bp from +97) was expressed as a GST fusion. Purified proteins were used to immunize guinea-pigs (Eurogentec). Antibodies were used in a 1:400 dilution, the specificity was verified after RNAi (Supplementary Fig. 1). For cell-type-specific mRNA isolation, we isolated 200 brain imaginal disc complexes for each experiment. Larvae expressing *UAS-PABP^{Flag}* (upstream activator sequence–poly(A)-binding protein gene–FLAG) specifically in neurons (*elav-GAL4*) or in glial cells (*repo-GAL4*) were dissected. PABP^{Flag} and adhering mRNA molecules were immunoprecipitated using anti-Flag antibodies. Following a mild fixation in 1% formaldehyde for 30 min mRNA was isolated as described[30,31]. In a subsequent RT–PCR reaction the following primers were used: *repo*, 5′-GAAGCCCGATGAGATGTGTT-3′, 5′-TAGTGAATGGTGGGGCTAGG-3′; *elav*, 5′-TCGTGCTTGTGTGCTCTTTC-3′, 5′-CCTGCTGTTGTTGCTGCT-3′; *pyramus*, 5′-CAACGTACAAGCCCATGTTG-3′, 5′-CACTCCTTTGTGGCGTTTCT-3′; and *thisbe*, 5′-CCGGCAGTAAATGGGTTAAA-3′, 5′-ACGGAACGGAACACGAAATA-3′. The annealing temperature was 59 °C, 35 cycles were used to detect the *repo* and *elav* gene products, 45 cycles were used to detect *pyramus* and *thisbe* complementary DNA.

Fly strains and genetics. All crosses were done on standard food at room temperature or 25 °C unless otherwise indicated. The following GAL4 and UAS stocks were obtained through the Bloomington stock centre. GAL4 driver strains: *repo-GAL4* (pan-glial), *Mz97-GAL4* (wrapping-glia), *elav-GAL4* (pan-neuronal), *eyeless-GAL4* (eye disc), *GMR-GAL4* and *LGMR-GAL4* (photoreceptor neurons). *UAS* effectors: *UAS-mCD8-GFP*, *UAS-ths* (ref. 14), *P(XP)d06722* (Exelixis Collection), *UAS-Ras85^{N17}*, *UAS-Ras^{N17}*, *UAS-htl^{DN}* and *UAS-λ-htl* (Bloomington).

For RNA interference studies, transgenic flies carrying a *UAS-htl* dsRNA were generated using standard procedures. A 421-bp fragment was amplified using the primers 5′-CGGACGTCTAGACATGGCGGAGGTCAATAAT-3′ and 5′-CCCTCGCCAGTCTAGAATCAGCAATCTTCAGC-3′. A *UAS-pyr* dsRNA construct was generated using the primers 5′-CACCTATGCAACTGGTTGAGCTGC-3′ and 5′-TCGCTGTTGTCTCAACTTGG-3′, no off-targets are predicted. Other dsRNA strains were obtained through the VDRC [transformant ID]: *ths^{dsRNA}* [24536]; *dof* [21317]; *sprouty^{dsRNA}* [6948]; *pointed^{dsRNA}* [7170]; *rolled^{dsRNA}* [43123]; *R^{dsRNA} (Rap1)* [33437] and *Ras85D^{dsRNA}* [18129] or through NIG-fly: *Ras64B^{dsRNA}* [1176R-1]. Other fly stocks used in this study: *repo-Flp5* (provided by M. Silies), *htl^{AB42}* (ref. 12), *FRT82B dof^{1}* (provided by M. Leptin), *FRT82B pnt^{A88}*, *P(PZ)eIF5B^{09143}*, *l(3)09143^{09143}*, *ths^{e02026}* and *pyr^{c02915a}* (Bloomington). *pyr^{c02915a}* is not a null allele[16], because *pyr^{c02915a}* / *Df(2R)BSC25* flies show an enhanced mutant glial phenotype. To drive GAL4-independent expression of GFP in glial cells we used *repo-lexA, lexAop-CD2-GFP*[32]. To express Pyramus we used the *pyr^{XP06722}* insertion.

Generation of clones. MARCM clones[13] were generated with *repo-Flp5; FRT82B tub-GAL80* and *repo4.3-GAL4, UAS-CD8-GFP FRT82B htl^{AB42} /TM6B, Tb, repo4.3-GAL4, UAS-CD8-GFP; FRT82B dof^{1} /TM6B, Tb* or *repo4.3-GAL4, UAS–CD8–GFP; FRT82B pnt^{A88} /TM6B, Tb*. Clonal patches of GAL4 expression were induced by *hsFlp; tubP>64bp>GAL4, UAS-GFP* (provided by M. Gonzalez-Gaitan). To induce recombination, 3-day-old larvae were transferred to 37 °C for 30 min.

31. Yang, Z., Edenberg, H. J. & Davis, R. L. Isolation of mRNA from specific tissues of *Drosophila* by mRNA tagging. *Nucleic Acids Res.* **33,** e148 (2005).
32. Lai, S. L. & Lee, T. Genetic mosaic with dual binary transcriptional systems in *Drosophila*. *Nature Neurosci.* **9,** 703–709 (2006).

©2009 Macmillan Publishers Limited. All rights reserved

Vol 460|6 August 2009|doi:10.1038/nature08184

LETTERS

Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*

Johan Malmström[1]*, Martin Beck[1]*, Alexander Schmidt[1,2], Vinzenz Lange[1,2], Eric W. Deutsch[3] & Ruedi Aebersold[1,2,3,4]

Mass-spectrometry-based methods for relative proteome quantification have broadly affected life science research. However, important research directions, particularly those involving mathematical modelling and simulation of biological processes, also critically depend on absolutely quantitative data—that is, knowledge of the concentration of the expressed proteins as a function of cellular state. Until now, absolute protein concentration measurements of a considerable fraction of the proteome (73%) have only been derived from genetically altered *Saccharomyces cerevisiae* cells[1], a technique that is not directly portable from yeast to other species. Here we present a mass-spectrometry-based strategy to determine the absolute quantity, that is, the average number of protein copies per cell in a cell population, for a large fraction of the proteome in genetically unperturbed cells. Applying the technology to the human pathogen *Leptospira interrogans*, a spirochete responsible for leptospirosis[2], we generated an absolute protein abundance scale for 83% of the mass-spectrometry-detectable proteome, from cells at different states. Taking advantage of the unique cellular dimensions of *L. interrogans*, we used cryo-electron tomography morphological measurements to verify, at the single-cell level, the average absolute abundance values of selected proteins determined by mass spectrometry on a population of cells. Because the strategy is relatively fast and applicable to any cell type, we expect that it will become a cornerstone of quantitative biology and systems biology.

The developed strategy combines three mass-spectrometry-based proteomic methods: absolute quantification using isotope-labelled reference peptides[3], label-free quantification, and high-throughput proteome sequencing by liquid chromatography–tandem mass spectrometry (LC–MS/MS)[4]. In the first step, we used isoelectric focusing by off-gel electrophoresis to fractionate tryptic digests of whole-cell protein extracts[5] and high-performance liquid chromatography–matrix-assisted laser desorption (HPLC–MALDI) and liquid chromatography–electrospray ionization (LC–ESI) tandem mass spectrometry using directed precursor ion selection[6] to identify the peptides contained in the respective fractions. Under the selected growth conditions we could identify 2,221 proteins, corresponding to 61% of the open-reading frames (ORFs) predicted from the *L. interrogans* genome[2]. From more than 90 LC–MS/MS runs, more than 410,000 fragment ion spectra (MS/MS spectra) were acquired, of which 145,703 were assigned to 18,303 unique peptides at a peptide false discovery rate (FDR) of less than 1%[7]. The identified peptides and proteins were assembled into a PeptideAtlas instance as previously described[8] (http://www.peptideatlas.org; see Supplementary Information for more information).

In the second step, we selected 32 peptides from the PeptideAtlas corresponding to 19 proteins at different abundance levels, determined by the number of matched MS/MS spectra acquired in the first step (spectral counts). The absolute abundance levels for the 19 proteins were determined using selected-reaction monitoring (SRM) and heavy-stable-isotope-labelled reference peptides[3] (Supplementary Table 1). By knowing the number of cells used to generate the sample and the amount of the heavy labelled peptides added, the copy number for the selected proteins could be calculated. The cellular abundance of these anchor proteins ranged from 40 to 15,000 copies per cell (Supplementary Table 1).

In the third step we used extracted precursor ion intensities for peptides derived from LC–MS maps acquired from trypsinized cell lysates. We then calculated the total protein ion intensity for the respective proteins by using the median intensities from the three most intense peptides matching to a specific protein[9–11]. The 19 anchor proteins with SRM-determined copy numbers then acted as calibration points for translating the relative abundance measurements based on extracted peptide precursor ion intensities[4] and spectral counting into absolute abundance measurements. Absolute protein abundance estimates were thereby obtained for 769 proteins using extracted ion intensities, and the absolute abundance of a further 1,095 proteins was estimated by spectral counting[12] (Supplementary Table 2). The number of proteins with estimated absolute protein abundance corresponds to 51% of the ORFs predicted from the *L. interrogans* genome. At a logarithmic scale, the extracted ion intensities correlate well with the absolute protein abundance (Fig. 1a). The accuracy of the absolute abundance measurement was determined by bootstrapping the extracted protein ion intensities against the SRM values. This statistical analysis allows different sets of reference peptides to be validated independently, by randomly removing a fraction of the data set, rebuilding the linear model and estimating the protein concentration of the initially removed data points. Because the real value of these data points is known from the SRM measurements, the average error can be estimated by multiple sampling events. The bootstrapping provided an estimated average error rate of 1.8-fold for the extracted ion intensities, and ~threefold for the spectral counting (Fig. 1b and Supplementary Fig. 4). Therefore, by using this three-step approach relying on three complementary mass spectrometry methods, we generated a proteome map of absolute protein concentrations for 51% of the ORFs predicted from the *L. interrogans* genome, corresponding to 83% of the proteome observable by deep proteome mapping, with an average error rate of less than threefold.

To assess the accuracy of the absolute protein abundance values generated by mass spectrometry, we applied cryo-electron tomography (cryoET) as an orthogonal and independent method. The extraordinarily thin cross section of *L. interrogans* cells (100–180 nm) makes

[1]Institute of Molecular Systems Biology, ETH Zurich (Swiss Federal Institute of Technology), Wolfgang Pauli-Strasse 16, CH-8093 Zurich, Switzerland. [2]Competence Center for Systems Physiology and Metabolic Diseases, CH-8093 Zurich, Switzerland. [3]Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA. [4]Faculty of Science, University of Zurich, CH-8057 Zurich, Switzerland.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

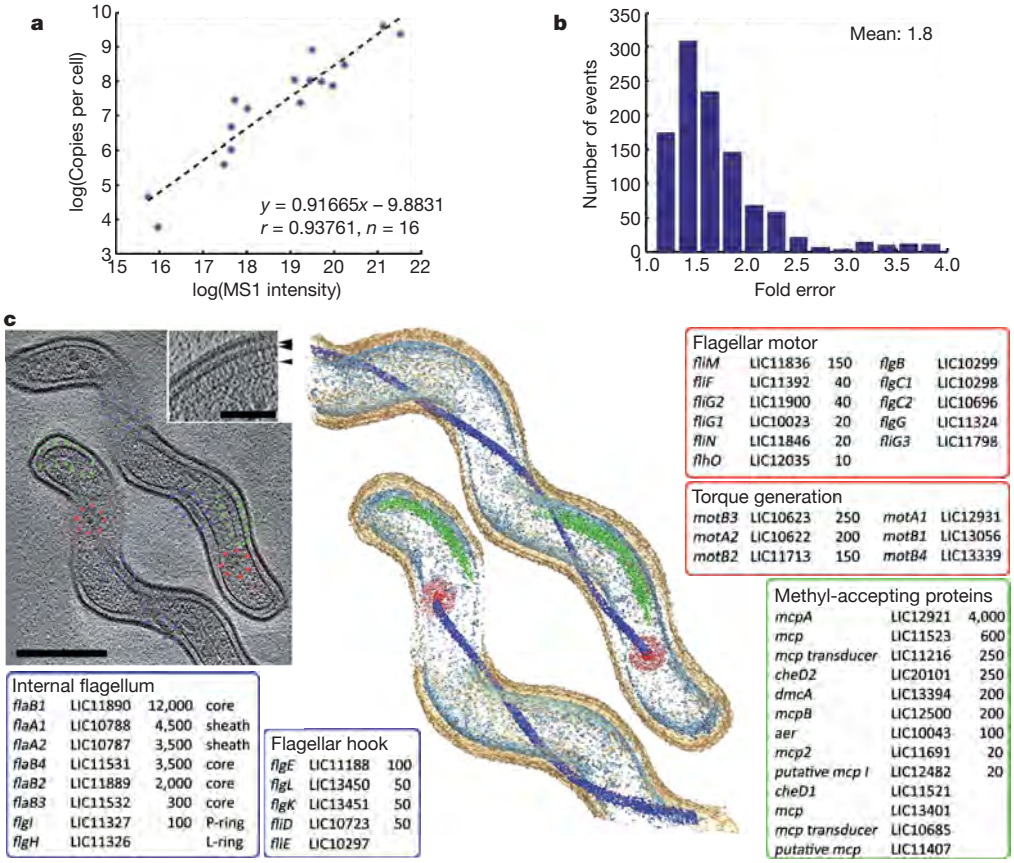


Figure 1 | Large-scale determination of cellular protein concentrations. **a**, Natural logarithm of extracted precursor ion intensities plotted against the natural logarithm of copies per cell for 16 proteins quantified by SRM. **b**, Distribution of error rates determined by bootstrapping. **c**, Slice through tomographic reconstruction, substructures are colour-coded as described below (scale bar, 200 nm). The inset shows a close-up of methyl-accepting proteins (scale bar, 100 nm). Arrowheads indicate the periplasmic MCP domain, plasma membrane and globular cytoplasmic MCP domain (from top). The boxes show the gene products making up the different components of methyl-accepting chemotaxis protein receptors (green), periplasmic flagella (dark blue), the flagellar stator (transparent red) and rotor (dark red). The accession numbers are from the NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez), and the values represent the numbers of copies per cell.

them an ideal specimen for cryoET measurements[13]. We assessed the accuracy of the mass-spectrometry-derived absolute abundance values *in vivo* by benchmarking against distinct morphological features with known subunit composition and structure, using a library of cryo-electron tomograms, covering subvolumes of more than 40 individual *L. interrogans* cells. The selected structures used to benchmark the absolute protein abundance estimations have been studied in depth both *in vitro* and *in vivo* using biochemical and structural biology techniques, and their structure and composition are known. The following features were analysed: (1) flagellar length, (2) flagellar motor, (3) methyl-accepting chemotaxis protein (MCP) receptors, and (4) total cellular protein concentration.

L. interrogans cells contain two periplasmic flagella that emanate from flagellar motors at both poles (Fig. 1c, red) and protrude towards the middle of the cell. The flagella are obvious features in *L. interrogans* cryo-electron tomograms (Fig. 1 and Supplementary Fig. 3, blue). FlaB1, the major core component and the most abundant flagellar protein, was estimated by mass spectrometry at 12,000 copies per cell. We determined that each cell contained 2,000 copies of flagellar protein FlaB2, 300 copies of FlaB3 and 3,500 copies of FlaB4. Both flagella combined thus contain an estimated average of 17,800 FlaB proteins organized into 11 protofilaments[14], with an intersubunit spacing of ~52 Å. From these data we calculated an average length of both flagella of 8.4 μm. This correlates well with the measured flagellar length, on the basis of the following considerations. The average cell length is 11.5 μm (determined from low magnification projection images, data not shown), and the combined length of both flagella is lower than the average cell length, as is apparent from the tomograms as shown in Supplementary Fig. 3.

The structure and protein composition of bacterial flagellar motors have been previously determined in various species[15,16]. The flagellar motor is an obvious structure in tomograms of *L. interrogans* (Fig. 1, red features). One of its components, FliF, has been shown to occur in 26 copies per motor in *Salmonella typhimurium*[15,16]. From the acquired tomograms we conclude that each *L. interrogans* cell has one motor at each pole, and assuming evolutionary conservation of the subunit composition, FliF is expected to occur in 52 copies per cell. For this low abundant protein, the average FliF copy number measured by mass spectrometry was 43, which is in agreement with the expected value within the estimated accuracy of the method.

MCPs function in metabolite sensing[17], forming arrays that are straightforward to discern by cryoET[18] and localize to the proximity of the flagellar motor in *L. interrogans* (Fig. 1c, green features). A receptor unit cell contains six molecules of MCPs that occupy an area of ~50 nm^2. By quantitative mass spectrometry we found 6,000 MCPs per cell, which translates into an estimated area of about 50,000 nm^2 per cell occupied by the receptors. This estimate is in agreement with the area occupied by receptor arrays in the tomograms, which was shown to occupy 40,000 nm^2 (200 nm × 100 nm per pole).

On the basis of the proteome-wide absolute abundance data and the cell volume measurements by cryoET, we determined the total cellular protein concentration to be ~250 mg ml^{-1} by summing all protein copies and dividing by the volume. This protein concentration is in agreement with earlier studies in *Escherichia coli*[19].

©2009 Macmillan Publishers Limited. All rights reserved

In summary, on the basis of the validation of the quantitative mass spectrometry data with an orthogonal method, we can confirm the estimated accuracy of the determined absolute abundance protein scale. The dynamic range of absolute abundance scale spans minimally three orders of magnitude from 40,000 copies per cell for protein LipL32 to single digit protein copies for low abundant proteins (Supplementary Table 2). The mass spectrometric approach described here is generic, whereas the validation of the data by cryoET is dependent on the cellular dimensions and is therefore not generic.

After the original proteome map of a bacterial species has been established, absolute protein quantification in repeat measurements of differentially perturbed cells is possible and straightforward. We investigated the quantitative changes in the *L. interrogans* proteome caused by treatment with ciprofloxacin, an antibiotic that inhibits DNA-gyrase[20]. *Leptospira interrogans* cells were counted and collected at three different time points of treatment, 3, 24 and 48 h, in five independent biological experiments, and the proteins were extracted and digested followed by mass spectrometric analysis. On average, around 1,000 proteins per state could be identified and quantified with absolute protein concentrations. Within the set of observed proteins, more than 200 proteins changed their protein abundance more than twofold (Supplementary Table 2).

In contrast to typical quantitative proteomics investigations, in which relative abundance levels are obtained that support the comparison of protein abundance between the same protein across samples, this approach allows comparisons between different proteins across samples. This is exemplified in Fig. 2, in which the identified proteins were grouped into biological functions using Gene Ontology analysis[21]. By taking into account the protein copy numbers determined in this study and the protein length, the fraction of the cellular protein synthesis budget associated with a particular cellular function could be calculated. The data indicate a considerable discrepancy between the number of ORFs and the number of protein molecules within a certain Gene Ontology group. More than 40% of the ORFs (30% of the identified proteins) in *L. interrogans* are hypothetical, and are associated with unknown biological processes. By summing up the total copy number of all proteins belonging to this group and multiplying by the protein length and comparing it to the total number of cellular proteins and their protein length, we estimate that the hypothetical proteins constitute only 12.7% of the total cellular protein synthesis budget (Fig. 2, blue). This indicates that the hypothetical proteins are generally of low abundance. *Leptospira interrogans* cells invest a large fraction of their cellular protein synthesis budget on the processes of protein synthesis and folding, cell motility and especially the proteins of the external encapsulating structure. In contrast, chemotaxis, even though it involves a large number of genes, has only a moderate impact on the protein synthesis budget. It is noteworthy that the cell invests a large proportion of the total protein synthesis budget to maintain a relatively small group of proteins at a very high cellular concentration, as in the case of the proteins of the external encapsulating structure (Fig. 2, bright green), suggesting that the functions carried out by these proteins are critical for the cell (for more information see Supplementary Information).

The ability to detect the absolute concentration of a significant fraction of the proteome also allowed us to examine how the proteome as a whole compensated for the changes of expression of specific proteins. As an example, after exposure to ciprofloxacin the cells reacted to the DNA-topoisomeric stress by a 15-fold up-regulation of recombinase A, a measurement that is consistent with literature values[22,23]. The increase in recombinase A was accompanied by an enormous increase in 15 hypothetical proteins that comprised approximately 20% of the entire proteome after ciprofloxacin treatment (Fig. 2). Interestingly, this large redistribution of the proteome did not significantly change the total cellular protein concentration. Therefore, the large increase in the abundance of the group of previously hypothetical proteins after ciprofloxacin exposure was compensated by a slight reduction of other high abundant protein classes, such as the ribosomal proteins and proteins involved in nucleotide binding (Fig. 2, red). This indicates that in *L. interrogans*, the cells strive to maintain a certain total number of protein components, that is, a constant cellular proteome concentration.

In this study, we describe an integrated mass spectrometric technique for the determination of the absolute concentration of proteins representing a significant fraction of the proteome of genetically unperturbed microbial cells. The technique was applied to the proteome of *L. interrogans*, a microbial species with a sequenced genome containing 3,658 predicted ORFs[2,24]. Out of the 2,221 identified proteins, 1,864 proteins, representing 51% of the predicted proteome, were provided with estimated copy per cell numbers. The cellular protein concentrations estimated by the technique were

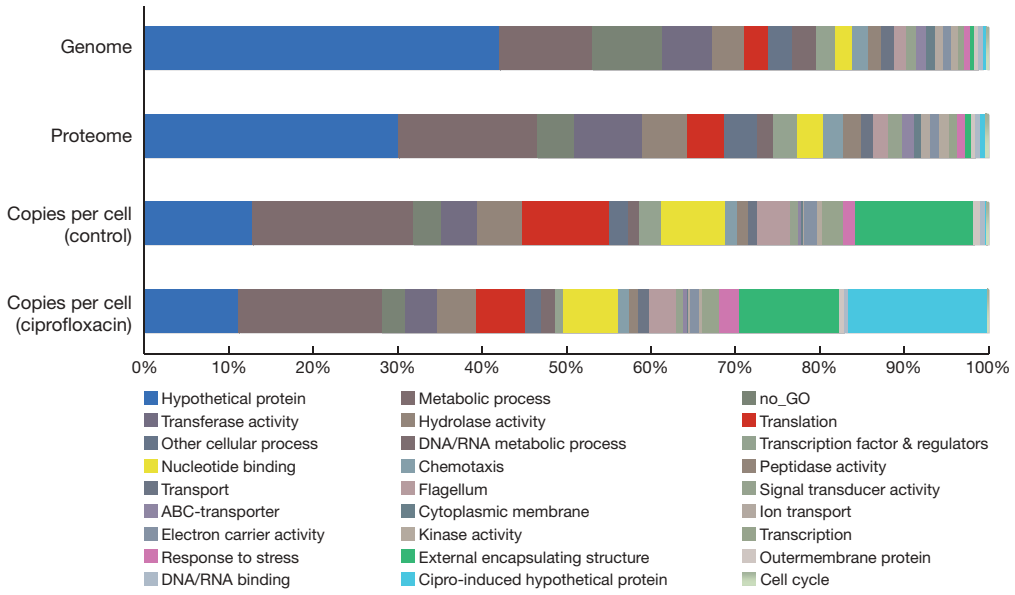


Figure 2 | Abundance levels of selected protein groups, by Gene Ontology analysis. The number of genes, the number of identified proteins and the copies per cell for two cellular states (control and ciprofloxacin treated). Please note that hypothetical proteins are underrepresented in the copies per cell calculation as compared to gene numbers, whereas members of the protein folding and encapsulating structure group are largely overrepresented. 'no_GO' denotes no Gene Ontology available.

©2009 Macmillan Publishers Limited. All rights reserved

assessed by bootstrapping and by the orthogonal method, cryoET tomography, which allowed us to detect, quantify and localize specific protein complexes in single *L. interrogans* cells in near-life state. The cryoET tomography measurements were possible on whole *L. interrogans* cells owing to the extraordinarily high cytoplasmic contrast and the unusual dimensions of the cells. The validated method was applied to study the reorganization of the *L. interrogans* proteome after exposure to the antibiotic ciprofloxacin. The data indicate that the cells react by expressing massive amounts of a small number of normally unexpressed proteins of unknown function while keeping the total cellular protein constant. The described technique is fast, efficient and can be applied to various biological systems of low and medium complexity in future studies.

METHODS SUMMARY

Leptospira interrogans serovar Copenhageni cells of the strain Fiocruz L1-130 were cultivated as previously described[25] and perturbed for 24 h with 5 μg ml^{-1} ciprofloxacin (antibiotic treatment). To establish the proteome map, peptides from trypsinized full cell lysates were separated by off-gel electrophoresis using twice the 3–10 pI range and once the 3–7 pI range. After isoelectric focusing, the fractions were subjected to LC–MS/MS as previously described[5,26]. MS/MS spectra were searched using SEQUEST against the predicted proteome from *Leptospira interrogans* serovar Copenhageni str, complete genome (NCBI accession numbers NC_005823 and NC_005824). The data were integrated into PeptideAtlas as previously described[8], and are available for browsing and downloading at http://www.peptideatlas.org/. To determine the absolute protein quantities, the cells were counted and collected by centrifugation. After cell lysis, the proteins were digested with trypsin followed by C18 reversed-phase peptide clean-up. Peptide quantification was determined by SRM as previously described[27], on the basis of heavy labelled reference peptides. For determination of MS1 (precursor ion) intensities, samples were analysed with a hybrid LTQ-FT-ICR mass spectrometer operated as previously described[6] and MS1-based peak extraction and alignment into MasterMaps were done using SuperHirn[4]. Cells from stimulated or non-stimulated cultures, respectively, were pipetted onto Quantifoil R2/1, 200 mesh copper grids (Plano), rapidly plunge-frozen into liquid ethane[28] and then introduced into a Technai F20 cryo-electron microscope (FEI). Tomograms were acquired and reconstructed as previously described[29].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 March; accepted 29 May 2009.
Published online 15 July 2009.

1. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
2. Ren, S. X. *et al.* Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature* **422,** 888–893 (2003).
3. Gerber, S. A. *et al.* Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl Acad. Sci. USA* **100,** 6940–6945 (2003).
4. Mueller, L. N. *et al. SuperHirn*—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7,** 3470–3480 (2007).
5. Malmstrom, J. *et al.* Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **5,** 2241–2249 (2006).
6. Schmidt, A. *et al.* An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol. Cell. Proteomics* **7,** 2138–2150 (2008).
7. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392 (2002).
8. Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Res.* **34** (database issue), D655–D658 (2006).
9. Silva, J. C. *et al.* Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144–156 (2006).
10. Vogel, C. & Marcotte, E. M. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nature Protocols* **3,** 1444–1451 (2008).
11. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272 (2005).
12. Lu, P. *et al.* Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnol.* **25,** 117–124 (2007).
13. Lucic, V., Forster, F. & Baumeister, W. Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.* **74,** 833–865 (2005).
14. Samatey, F. A. *et al.* Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature* **410,** 331–337 (2001).
15. Jones, C. J., Macnab, R. M., Okino, H. & Aizawa, S. Stoichiometric analysis of the flagellar hook-(basal-body) complex of *Salmonella typhimurium*. *J. Mol. Biol.* **212,** 377–387 (1990).
16. Sosinsky, G. E. *et al.* Mass determination and estimation of subunit stoichiometry of the bacterial hook-basal body flagellar complex of *Salmonella typhimurium* by scanning transmission electron microscopy. *Proc. Natl Acad. Sci. USA* **89,** 4801–4805 (1992).
17. Charon, N. W. & Goldstein, S. F. Genetics of motility and chemotaxis of a fascinating group of bacteria: the spirochetes. *Annu. Rev. Genet.* **36,** 47–73 (2002).
18. Briegel, A. *et al.* Location and architecture of the *Caulobacter crescentus* chemoreceptor array. *Mol. Microbiol.* **69,** 30–41 (2008).
19. Elowitz, M. B. *et al.* Protein mobility in the cytoplasm of *Escherichia coli*. *J. Bacteriol.* **181,** 197–203 (1999).
20. Shalit, I., Barnea, A. & Shahar, A. Efficacy of ciprofloxacin against *Leptospira interrogans* serogroup icterohaemorrhagiae. *Antimicrob. Agents Chemother.* **33,** 788–789 (1989).
21. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
22. Dwyer, D. J., Kohanski, M. A., Hayete, B. & Collins, J. J. Gyrase inhibitors induce an oxidative damage cellular death pathway in *Escherichia coli*. *Mol. Syst. Biol.* **3,** 91 (2007).
23. Reckinger, A. R., Jeong, K. S., Khodursky, A. B. & Hiasa, H. RecA can stimulate the relaxation activity of topoisomerase I: Molecular basis of topoisomerase-mediated genome-wide transcriptional responses in *Escherichia coli*. *Nucleic Acids Res.* **35,** 79–86 (2007).
24. Nascimento, A. L. *et al.* Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. *J. Bacteriol.* **186,** 2164–2172 (2004).
25. Haake, D. A. *et al.* Changes in the surface of *Leptospira interrogans* serovar grippotyphosa during *in vitro* cultivation. *Infect. Immun.* **59,** 1131–1140 (1991).
26. Yi, E. C., Lee, H., Aebersold, R. & Goodlett, D. R. A microcapillary trap cartridge-microcapillary high-performance liquid chromatography electrospray ionization emitter device capable of peptide tandem mass spectrometry at the attomole level on an ion trap mass spectrometer with automated routine operation. *Rapid Commun. Mass Spectrom.* **17,** 2093–2098 (2003).
27. Lange, V. *et al.* Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. *Mol. Cell. Proteomics* **7,** 1489–1500 (2008).
28. Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.* **21,** 129–228 (1988).
29. Beck, M. *et al.* Snapshots of nuclear pore complexes in action captured by cryo-electron tomography. *Nature* **449,** 611–615 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project has been funded in part by ETH Zurich, the Swiss National Science Foundation (grant 31000-10767), federal funds from the National Heart, Lung and Blood Institute, the National Institutes of Health (contract no. N01-HV-28179), SystemsX.ch, the Swiss initiative for systems biology, in part by the PROSPECTS (proteomics in time and space) European network of excellence, and with funds from the ERC project 'Proteomics V3.0' for R.A. J.M. was supported by a fellowship from the Swedish Society for Medical Research (SSMF), M.B. was supported by a long-term fellowship of the European Molecular Biology Organization and a Marie Curie fellowship of the European Commission, A.S. and V.L. were supported by the Competence Center for Systems Physiology and Metabolic Diseases. We thank O. Medalia and the electron microscopy facility of ETH Zurich (EMEZ) for support, and D. A. Haake for critical reading of the manuscript.

Author Contributions J.M. and M.B. planned the experiments, performed the experimental work and data analysis and wrote the manuscript. A.S. and V.L. participated in the experimental work and the data analysis and E.W.D. assembled the PeptideAtlas build. R.A. was the project leader and wrote the manuscript.

Author Information The mass spectrometry data, including spectra and the identified peptides and proteins, have been deposited into a PeptideAtlas instance found at http://www.peptideatlas.org/builds/ (*Leptospira interrogans*, Jan 2008 Build). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.A. (aebersold@imsb.biol.ethz.ch).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Cell culture and treatment. The *Leptospira interrogans* serovar Copenhageni of the strain Fiocruz L1-130 were obtained from the American Type Culture Collection (ATCC) and cultivated as described previously[25]. Cultures of 30 ml were grown at 30 °C to a density of 2×10^7 cells ml^{-1}, and treated with the antibiotic ciprofloxacin (5 μg ml^{-1}) for 3, 24 and 48 h. Afterwards, the cells were collected by centrifugation at 3,000*g*, washed once in PBS, counted, pelleted again, resuspended in 2 ml denaturation buffer (100 mM HEPES, pH 7.6, 6 M urea), sonicated for 5 min and stored at −80 °C.

Protein digestion. The proteins were reduced with 5 mM dithiothreitol (DTT) for 45 min at 37 °C, and alkylated with 25 mM iodoacetamide for 45 min in the dark before diluting the sample with 100 mM HEPES at pH 8.5 to a final urea concentration below 1.5 M urea. Proteins were digested by incubation with trypsin (1/100, w/w) for at least 6 h at 37 °C. The peptides were cleaned up by C18 reversed-phase spin columns according to the manufacturer's instructions (Harvard Apparatus).

Off-gel electrophoresis. The dried-down peptides were resolubilized to a final concentration of 1 mg ml^{-1} in off-gel electrophoresis buffer containing 6.25% glycerol and 1.25% IPG buffer (GE Healthcare). The peptides were separated on both pH 3–10 IPG strips and pH 3–7 IPG strips (GE Healthcare) with a 3100 OFFGEL fractionator (Agilent) as previously described[30], using a protocol of 1 h rehydration at maximum 500 V, 50 μA and 200 mW followed by the separation at maximum 8,000 V, 100 μA and 300 mW until 50 kVh were reached. After isoelectric focusing, the fractions were concentrated and cleaned up by C18 reversed-phase spin columns according to the manufacturer's instructions (Harvard Apparatus).

Shotgun mass spectrometry. The setup of the micro-capillary reversed phase liquid chromatography (μRPLC)-MS system was as described previously[5,26]. ESI-based LC–MS/MS (LTQ ThermoFinnigan, Thermo Fisher Scientific) analyses were carried out using an Agilent 1100 series (Agilent Technologies) on a 75 μM × 10.5 cm fused silica microcapillary reversed phase column. After sample loading, the samples were separated by a 65 min linear gradient of 5 to 35% acetonitrile in water, containing 0.1% formic acid, with a flow rate of 200 nl min^{-1}. Peptides eluting from the capillary column were selected for collision activated dissociation (CAD) by the mass spectrometer using a protocol that alternated between one MS scan and three MS/MS scans. The specific *m*/*z* value of the peptide fragmented by CAD was excluded from reanalysis for 2 min using the dynamic exclusion option.

The hybrid LTQ-FT-ICR mass spectrometer was interfaced to a nanoelectrospray ion source (both from Thermo Electron) coupled online to a Tempo 1D-plus nanoLC (Applied Biosystems/MDS Sciex). Peptides were separated on a RPLC column (75 μm × 15 cm) packed in-house with C18 resin (Magic C18 AQ 3 μm; Michrom BioResources) using a linear gradient from 98% solvent A (98% water, 2% acetonitrile, 0.15% formic acid) and 2% solvent B (98% acetonitrile, 2% water, 0.15% formic acid) to 30% solvent B over 90 min at a flow rate of 0.3 μl min^{-1}. Three MS/MS spectra were acquired in the linear ion trap per each Fourier-transform ion cyclotron resonance (FT-ICR)-MS scan which was acquired at 100,000 full-width at half-maximum nominal resolution settings with an overall cycle time of approximately 1 s. The specific *m*/*z* value of the peptide fragmented by CAD was excluded from reanalysis for 0.5 min using the dynamic exclusion option. Charge state screening was used to select for ions with at least two charges and reject ions with undetermined charge state. The normalized collision energy was set to 32%, and one microscan was acquired for each spectrum. For quantification, the SuperHirn peak extraction and alignment algorithm were used and peptide quantities were determined from the ion current of the specific signal[4].

HPLC and spotting for MALDI analysis. The different sample fractions were loaded onto a RP capillary column (100 μm i.d. × 15 cm length) by a Famos micro-autosampler from Eksigent. The column was in-house packed with Magic C18AQ (200 Å pore, 5 μm diameter, Michrom Bioresources) onto the capillary tubing with a borosilicate frit (Integrafrit, New Objective). Chromatographic separations were carried out with a nanoLC pump (Eksigent) at 500 nl min^{-1} flow rate using a solvent composition gradient of solvent A (water/acetonitrile/trifluoroacetic acid, 98/2/0.1, v/v/v) and B (water/acetonitrile/trifluoroacetic acid, 2/98/0.1, v/v/v). A linear binary gradient of 5–40% solvent B was generated over 50 min, followed by 5 min flush at 90% solvent B. The eluting peptides were mixed with MALDI matrix (3 mg ml^{-1} of α-cyano-4-hydroxycinnamic acid in 70% acetonitrile and 0.1% trifluoroacetic acid spiked with angiotensin II, 0.5 pmol ml^{-1} (Sigma) and ACTH, 1.25 pmol ml^{-1} (Proteomass)) delivered with a flow rate of 1.4 μl min^{-1} and spotted on to the MALDI targets at a 5-s interval by MALDI spotter of Tempo nano LC system (Applied Biosystems) for a total of 616 spots per gradient run.

MALDI–time of flight (TOF)/TOF mass spectrometry analysis. MS and MS/MS analysis was performed using a 4800 MALDI–TOF/TOF Analyser (Applied Biosystems). Each spot was first analysed in mass spectrometry mode, by accumulating signal with up to 1,000 laser shots (20 sub-spectra of 50 shots) over the mass range 800–4,000 Da unless a preset stop criteria of 10 sub-spectra was reached at which the accumulated spectrum contained at least five peaks with signal-to-noise (S/N) > 100. Up to ten ions of each spot giving a mass spectrometry signal with $S/N > 30$ were then candidates for further MS/MS analysis, performed in order of increasing precursor intensity. The job-wide interpretation, which generated the list of precursor ions and assigned the most intensive spot of a precursor ion for MS/MS analysis, was used for each sample, and the acquisition of an MS/MS spectrum was obtained by accumulating 1,500 laser shots (30 sub-spectra of 50 shots) with the collision energy of 1 kV. The source air pressure was set to 2.5×10^{-6} torr for MS/MS analysis and 5×10^{-7} torr for MS analysis.

Data processing and compilation of PeptideAtlas. MS/MS spectra were searched using the SEQUEST search tool[31] against the predicted proteome from *Leptospira interrogans* serovar Copenhageni str, complete genome NCBI genome number NC_005823 and NC_005824 (http://www.ncbi.nlm.nih.gov/entrez), consisting of 3,658 proteins, as well as known contaminants, such as porcine trypsin and human keratins (Non-Redundant Protein Database, National Cancer Institute Advanced Biomedical Computing Center, 2004, ftp://ftp.ncifcrf.gov/pub/nonredun/). The search was performed with semi-tryptic cleavage specificity, mass tolerance of 3 Da, methionine oxidation as variable modification and cysteine carbamidomethylation as fixed modification. The database search results were further processed using the PeptideProphet program[7] modified to include the pI information[5]. The PeptideAtlas build was constructed using the sequence search results of 145,703 MS/MS spectra with a $P \geq 0.9$. We identified 18,303 unique peptides at a peptide FDR of less than 1%[7], which represented 2,221 proteins at a protein FDR of less than 1%. Overall, the identified proteins matched to 61% of the predicted gene models and contained 75% (1,291 proteins) of the known or characterized proteins, 51% (85) of the proteins with putative function, and 49% (867) of the hypothetical proteins. The data was compiled into a publicly accessible instance of PeptideAtlas[8] to facilitate further studies of Leptospirosis using targeted proteomics. We used Gene Ontology analysis[21] to group the identified proteins according to their cellular function. All the data are available for browsing and downloading at http://www.peptideatlas.org/.

Targeted mass spectrometry. Absolute protein concentration was determined by SRM on the basis of heavy-labelled reference peptides that served as internal standards (Supplementary Table 1). Reference peptides were obtained from Thermo Fisher Scientific and Sigma-Aldrich in defined concentrations calibrated by amino acid analysis. SRM was performed on a hybrid quadrupole-linear ion trap mass spectrometer (4000 QTRAP) as previously described[27]. The instrument was coupled to a Tempo Nano LC system (Applied Biosystems/MDS Sciex) for peptide separation using a 30 min gradient from 5 to 30% acetonitrile (0.1% formic acid) at 300 nl min^{-1} flow rate. A fused silica emitter of 75 μm inner diameter was packed in-house with 15 cm Magic C18AQ (200 Å pore, 5 μm diameter, Michrom Bioresources). Quantitative analyses in SRM mode were performed with Q1 and Q3 operated in unit resolution (0.7 *m*/*z* half-maximum peak width). Five biological replicate measurements were carried out for each sample. The results from the absolute quantification experiments by SRM are summarized in Supplementary Table 1.

Cryo-electron tomography and image processing. Cells from stimulated or untreated cultures, respectively, were pipetted onto Quantifoil R2/1, 200 mesh copper grids (Plano), rapidly plunge-frozen into liquid ethane as described[28] and then introduced into a Technai F20 cryo-electron microscope (FEI) equipped with Gatan image filter. Tomograms were acquired as described[29] with the following modifications: an underfocus of 6.5 μm was used and the nominal magnification was ×34,000, corresponding to an object pixel size of 0.63 nm at the specimen level. All image processing operations were performed with the EM-package[32] or Tom-package[33] for Matlab (The MathWorks). The three-dimensional surface-rendered representations were created with Amira (TGS). Tomograms were initially reconstructed with a binning factor of two, as described earlier[29]. To calculate the average cell volume the subvolume of several cells covered by tomograms was extrapolated to the entire cell length measured in low magnification projection images.

30. Heller, M. *et al.* Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.* **4**, 2273–2282 (2005).
31. Eng, J. K., McCormack, A. L. & Yates, J. R. III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
32. Hegerl, R. The EM program package: a platform for image processing in biological electron microscopy. *J. Struct. Biol.* **116**, 30–34 (1996).
33. Nickell, S. *et al.* TOM software toolbox: acquisition and analysis for electron tomography. *J. Struct. Biol.* **149**, 227–234 (2005).

©2009 Macmillan Publishers Limited. All rights reserved

METHODS

Genome-wide genotyping. SGENE was initially made up of 1,321 cases and 12,277 controls typed at deCODE Genetics using the Illumina HumanHap300 BeadChip. For SGENE-plus, an additional 859 cases and 854 controls typed at Duke University using either the Illumina HumanHap300 BeadChip or the Illumina HumanHap550 BeadChip as well as 483 cases and 367 controls typed at the University of Bonn using the HumanHap550 BeadChip were also included. Samples were excluded if they were low yield (low yield was defined as below 98% except for the samples typed at Duke, in which case low yield was below 96%), if they were duplicates of other samples included in the study, if they had a sex determined by X chromosome marker homozygosity different from their reported sex or if they were estimated to have less than 90% European ancestry by running STRUCTURE[29] using the HapMap CEU, YRI and CHB/JPT individuals as reference samples. Of the 317,503 markers on the HumanHap300 BeadChip, 2,635 were deemed unusable due to lack of polymorphism, severe deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-10}$), low yield (<95% in either cases or controls) or allele frequency differences between the typing centres ($P < 1 \times 10^{-7}$); 314,868 markers, then, remained for analysis.

Follow-up genotyping. Follow-up set 1 (715 cases; 3,634 controls) was genotyped at UCLA on the HumanHap550 BeadChip and at deCODE Genetics on the HumanCNV370 BeadChip. Only the markers shown in Supplementary Table 3, however, were used in this study. Follow-up set 2 (3,330 cases; 6,892 controls) was genotyped at deCODE Genetics using Centaurus assays (Nanogen). Assay quality was evaluated by genotyping the CEU HapMap samples and comparing the results with the publicly released HapMap data. Assays with a greater than 1.5% mismatch rate were not used. Follow-up set 3 (287 cases; 3,987 controls) was typed in Finland using the Sequenom MassArray iPLEX genotyping system, following the manufacturer's instructions (Sequenom Inc.). Briefly, the system involves multiplex PCR and mini-sequencing assays, followed by MALDI-TOF mass spectrometry analysis. Follow-up set 4 (667 cases; 1,042 controls) was typed at the Santiago de Compostela node of the Spanish National Genotyping Centre (http://www.cegen.org) using the Sequenom MassArray iPLEX genotyping system, following the manufacturer's instructions (Sequenom Inc.). As a quality check, all clusters were manually inspected for accurate genotype assignment. In addition, 1,781 genotypes were successfully assayed twice, with no discordant results.

Association analysis. Association analysis was carried out using a likelihood procedure described in a previous publication implemented in NEMO software[26]. Allele-specific ORs and associated *P* values were calculated assuming a multiplicative model for the two chromosomes of an individual. Association was tested using a standard likelihood ratio statistic that, if the subjects were unrelated, would have asymptotically a chi-squared distribution with 1 degree of freedom under the null hypothesis. To correct for relatedness and potential population stratification in the genome-wide typed samples (SGENE-plus and follow-up set 1), genomic control was used[27]. Inflation factors, estimated by dividing the median of the 314,868 chi-squared statistics by 0.675^2, were 1.01, 1.03, 1.09, 1.05, 1.05, 1.19, 1.04 and 1.09 for the SGENE-plus England, Finland (Helsinki), Finland (Kuusamo), Germany (Bonn), Germany (Munich), Iceland, Italy and Scotland groups, respectively, and 1.08 for follow-up set 1. The inflation factor was large in Iceland because of the inclusion of close relatives in that data set. Both SGENE-plus and the follow-up samples were combined using the Mantel–Haenszel model[28].

Summary statistic combination. Combined *P* values for the SGENE-plus, International Schizophrenia Consortium and Molecular Genetics of Schizophrenia studies were calculated by summing *Z* scores with each group's *Z* score multiplied by the inverse of that group's standard error divided by the square root of the sum of the squared inverse standard errors. Combined ORs were calculated by summing log ORs with each log OR weighted by the inverse of its variance.

©2009 Macmillan Publishers Limited. All rights reserved

CORRIGENDUM

doi:10.1038/nature08286

Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year

Eric J. Steig, David P. Schneider, Scott D. Rutherford, Michael E. Mann, Josefino C. Comiso & Drew T. Shindell

Nature **457**, 459–462 (2009)

In this Letter, we reported trends on reconstructed temperature histories for different areas of the Antarctic continent. The confidence levels on the trends, as given in the text, did not take into account the reduced degrees of freedom in the time series due to autocorrelation. We report in Table 1 the corrected values, based on a two-tailed *t*-test, with the number of degrees of freedom adjusted for autocorrelation, using $N_{\text{effective}} = N(1 - r)/(1 + r)$, in which N is the sample size and r is the lag-1 autocorrelation coefficient of the residuals of the detrended time series. The median of r is 0.27, resulting in a reduction in the degrees of freedom from $N = 600$ to $N_{\text{effective}} = 345$ for the monthly time series.

We also include results of a further calculation that takes into account both the variance and the uncertainty in the reconstructed temperatures. We performed Monte-Carlo simulations of the reconstructed temperatures using a Gaussian distribution with variance equal to the unresolved variance from the split calibration/verification tests described in the paper. Confidence bounds were obtained by detrending each simulation and obtaining the lag-1 autocorrelation coefficient and variance of the residuals; a random realization of Gaussian noise having the same lag-1 autocorrelation coefficient and variance was then added to the trend, and a new trend was calculated. The 2.5th and 97.5th percentiles of the 10,000 simulated trends give the 95% confidence bounds. For the case of zero unresolved variance, this calculation converges on the same value as the two-tailed *t*-test, above. The 95% confidence minimum trend value is given by the 5th percentile values of the simulated trends, last row of Table 1.

The corrected confidence levels do not change the assessed significance of trends, nor any of the primary conclusions of the paper. We also note that there is a typographical error in Supplementary Table 1: the correct location of Automatic Weather Station 'Harry' is 83.0° S, 238.6° E. The position of this station on the maps in the paper is correct.

Table 1 | Corrected confidence levels on mean decadal temperature trends

	West Antarctica	East Antarctica	Antarctic Peninsula	All Antarctica
Trend (°C per decade)	0.18	0.10	0.11	0.12
95% CI of trend in mean reconstruction	±0.09	±0.10	±0.05	±0.10
95% CI of trend, accounting for unresolved variance in mean reconstruction	±0.12	±0.13	±0.07	±0.12
Minimum trend (95% confidence, accounting for unresolved variance in mean reconstruction)	0.08	−0.01	0.05	0.02

The confidence levels are shown over the period 1957–2006 for the reported surface temperatures based on satellite data. CI, confidence interval.

©2009 Macmillan Publishers Limited. All rights reserved

PROSPECTS

Back to first principles

We need more physician innovators not just more physician scientists, writes Justin Chakma.

Pre-med and medical-school curricula have not kept apace with advances in scientific knowledge, according to a report by the Howard Hughes Medical Institute and a committee of the Association of American Medical Colleges. Physicians need more basic-science training, the authors argue.

IMAGES.COM/CORBIS

But this is only half the story. Medical schools should not simply produce clinicians who understand the science of disease mechanisms. The more urgent need is for physicians who can translate research findings to a clinical setting — and most physician-scientist (MD/PhD) programmes do not focus adequately on this area.

MD/PhD courses typically generate physician-scientists who either become pure clinicians or focus on basic sciences like any PhD-trained biomedical scientist. But physician-scientists should be trained to translate innovations in the lab to the hospital.

Several medical schools are revamping traditional clinical-investigator programmes in an interdisciplinary, team-based approach to training. For example, Stanford School of Medicine, California, offers a graduate-level course, and the University of Michigan at Ann Arbor runs a one-year fellowship in medical innovation targeted at medical residents and MD/PhD students.

In our programme at the University of Toronto, Canada, participants choose a clinical speciality outside their area of interest. Team members learn a speciality from practising physician mentors and spend eight weeks in a hospital observing medical procedures. Participants then focus on just a few projects based on criteria such as the importance of the market, team preference and disease prevalence. Teams often reach animal trials and even human studies if further development is commercially viable. Students work with biomedical engineers, entrepreneurs, and others to make advances in surgery, imaging and regenerative medicine.

Physicians with strong basic science can play a facilitating role in reconciling the science behind prototypes and assays with the disease mechanisms underlying the clinical needs. They need not be the ones discovering — that can be left to the scientists.

These programmes offer a systematic, science-based approach to innovation. Simple collaboration means a more efficient deployment of scientifically literate physicians. What we need are not more physician-scientists, but more physician-innovators and physician-facilitators. ■

Justin Chakma is founder of BioDesign Toronto at the University of Toronto, Canada.

POSTDOC JOURNAL

Footloose and freelance?

I have decided to opt out of academia, at least for the foreseeable future. This will come as no surprise to those who know me, nor to anyone who has been following my Postdoc Journal. Although I feel that a great weight has been lifted from my shoulders, this decision raises another concern — my husband's postdoc salary alone won't support the family.

I am now entertaining the idea of becoming a freelance writer; the autonomy and ability to work from home hold great appeal. After dabbling with writing last year, I had some modest success and published articles on science and the environment in a number of magazines — although I wonder if this was beginner's luck. Finding assignments has become harder of late. Initially, I rather naively, and perhaps somewhat arrogantly, assumed that editors would be clamouring for articles from scientists with a PhD. Now I realize I am one of a legion of academics-turned-writers — a rookie in a world of shrinking magazine markets.

Still, like most academics, I am well-acquainted with rejection and my thick skin will surely serve me well as a freelancer. While I strive to rid myself of the terrible writing habits common to scientists, I have managed to find some writing and editing work from contacts in Australia and I hope to find assignments in the United States. I remain quietly optimistic about my freelancing future. ■

Joanne Isaac was a postdoc in climate-change effects on biodiversity at James Cook University, Townsville, Australia. She is now in the United States so that her husband can complete a postdoc.

IN BRIEF

Postdocs join union

Some 350 postdocs at Rutgers University in New Jersey have elected to join the union that already represents more than 5,000 faculty members and graduate employees at the university. The New Jersey Public Employment Relations Commission certified the vote late last month and the group will be represented by the Rutgers council of the American Association of University Professors–American Federation of Teachers. Chemistry department postdoc Alan Wan says postdocs' terms of employment, including compensation and benefits, had never been spelled out and were determined by principal investigators.

Help for service economy

Science, technology, engineering and maths can provide significant benefits to the United Kingdom's service sector but their role is hidden and unacknowledged, laments a report by the Royal Society. Such benefits could include much more accurate financial-risk reports for the financial services sector, according to *Hidden Wealth: The Contribution of Science to Service Sector Innovation*. The report makes some specific recommendations. One is for banking, technology and research groups to create systemic risk modelling and risk-assessment analyses using the latest research (see page 680). Another is for a formal cooperative exchange between research academics and the service sector.

Spotlight on UK energy

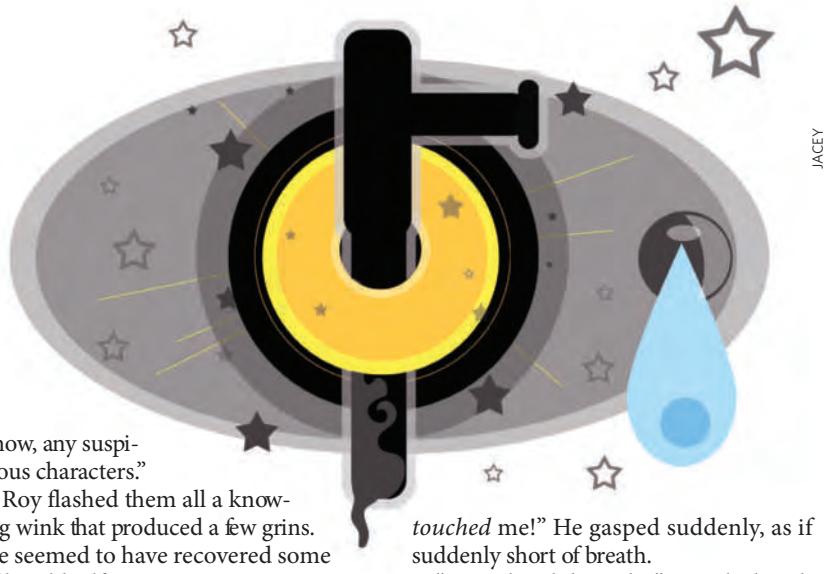
The UK Engineering and Physical Sciences Research Council is recruiting an international panel of scientists from academia, industry and other sectors to assess the country's energy-research programme. This will be the first ever review of energy research being carried out with funding from the UK government's research councils. Among the areas to be assessed are renewable and conventional energy sources, sustainable energy, and energy-demand reduction. The panel will conduct its review next April, and will present its results to the UK research community and the seven research councils. The assessment is part of the councils' oversight of the disciplines in which they fund research.

© 2009 Macmillan Publishers Limited. All rights reserved

Expatriate

Contact has been made.

Julian Tang



JACEY

Roy Gredenski grinned as his rookies roared with laughter at his latest tale. He was celebrating his thirtieth year in the Customs and Immigration Department.

"Roy, do you have any other stories for the youngsters?" grinned his captain, Joe Werner, from the back of the room, where his other senior colleagues were sitting. They'd heard them all before but the tales only seemed to get better with each retelling.

Roy paused, looking around at the young, eager faces surrounding him.

"Well, there is one story," he said quietly and paused.

He looked down at the floor, as if trying to decide how — or whether — to tell it. The room went quiet, his audience waiting, respectfully. Then, amazingly, Roy seemed to be on the verge of weeping.

"What is it Roy?" asked Joe, coming forward with some surprise and concern. There were soft sounds of sobbing coming from Roy by now.

"God, I didn't know," he gasped. "How would anyone have known?" he whispered.

Joe quickly ushered all the cadets out of the room. His remaining senior colleagues brought their chairs closer to Roy then sat and waited. After a few minutes, Roy raised his eyes and looked around at them, gratefully.

"My dear friends," he said softly. "After I tell you this, you won't want to know me."

"Go on, Roy," said Joe, who was sitting beside him, patting him gently on the back. "What is it?"

"Joe, you remember the first Moon Base Spaceport?"

Joe nodded.

"Well, you and I were both rookies the year it opened? Remember?"

"Yeah, I remember. What chaos!" he laughed. "Will, you were also there with us, weren't you?"

Will Devine, the second-in-command, nodded. "Yeah, that was some opening. We had lines that seemed to stretch all the way back to Earth, waiting to clear immigration. There were all sorts — businessmen, scientists, tourists, school kids — all wanting a peek at the Moon. We should have sold tickets!"

"Yeah, well, for the rest of you who weren't there, Joe and Will were out front with the travellers. I was in the back, in the hot-room — where we dealt with, you know, any suspicious characters."

Roy flashed them all a knowing wink that produced a few grins. He seemed to have recovered some of his old self.

"Do either of you remember that guy you brought in to me that day? You know, that tall, skinny guy with a funny face? You were both annoyed that he didn't have any papers."

"Yeah, that was the damndest thing. Not only did he not have any travel documents, he looked like something that had just stepped out of one of those old Frankenstein movies — like plastic surgery gone wrong," said Joe with feeling.

Will nodded vigorously in agreement. "Actually, we were really just mad because he just wasn't aware of the trouble he was causing for us. He seemed to have no idea that he needed some sort of travel document to get into the base. He was just like a kid — but didn't look like one," he finished with a shudder.

Roy sighed. "Well, whatever the reason, you guys brought him to me. That's where it started."

He paused a little before going on.

"You guys were right. He looked like someone had put him together wrong. In fact he was so weird I did a strip search on him — not because I thought he was really carrying any drugs, but because I just wanted to see the rest of him."

"And?" prompted Joe, curious.

"God, he really was a Frankenstein's monster! His ears and eyes weren't even at the same level. He actually had two left feet — and two right hands. And, man, he had no anus! There was just a dimple where it was supposed to be! I felt so sick I wanted to vomit."

He paused again.

"What happened next Roy?" asked Will.

"Well, I'm sure any of you would have reacted the same way," said Roy defensively, glancing around quickly as if for confirmation. "That thing *touched* me!" He gasped suddenly, as if suddenly short of breath.

"Roy, what did you do?" Joe asked, with sudden dread.

"I lost it, guys. His touch was so cold and … well, before I knew it I had hit him with my riot stick. It was just a reflex reaction."

There was a collective intake of breath.

"Roy, you told us that you sent him through after some further questioning — you said you were satisfied with his answers," Joe said in disbelief.

Roy seemed not to have heard him.

"He didn't bleed — well not blood anyway. His body just seemed to burst with that first strike, like a waterbed. This grey liquid seemed to come out of his eyes and nose — like he was leaking. After that, I did vomit. By the time I had got back to clean up the mess, his body seemed to have dissolved in my vomit."

There was another short silence then Roy looked around at them.

"When I searched his bag, you know what I found? You'll never believe this, so I'll show you."

He went over to his locker, unlocked it and took out an old envelope. He slid out what was inside and held it up for the group to see.

There was another collective gasp from the group.

"Isn't that …? Yeah, it is! I remember seeing it in books when I was at school. It's the welcome disc from that Voyager spaceship that was sent out hundreds of years ago!" said Joe, in awe.

"Yeah, it is," said Roy, starting to weep again.

This time, no one tried to comfort him.■

Julian Tang is a clinical/academic virologist. He would like to dedicate this story to his wife, Florence, who keeps him calm and sane, while navigating the delays and frustrations of modern international travel.

FUTURES

© 2009 Macmillan Publishers Limited. All rights reserved